UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

A METHOD FOR CALIBRATING PROBABILISTIC FORECASTS

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

PATRICK TIMOTHY MARSH
Norman, Oklahoma
2013

A METHOD FOR CALIBRATING PROBABILISTIC FORECASTS

A DISSERTATION APPROVED FOR THE
SCHOOL OF METEOROLOGY

BY

Dr. Kevin A. Kloesel, Co-Chair

Dr. John S. Kain, Co-Chair

Dr. David J. Stensrud

Dr. Michael B. Richman

Dr. Frederick H. Carr

Dr. S. Lakshmivarahan

Contents

List of Tables

List of Figures

Abstract

Chapter 1

Introduction

Rare meteorological events[1] that occur on small spatial and short temporal scales pose significant challenges to forecasters. This is related to the limited predictability of phenomena occurring on short time-space scales; however, these events comprise a substantial portion of meteorological phenomena that negatively impact society (e.g., heavy rain, large hail, tornadoes, etc.). Thus, "good" forecasts of these events would provide large societal benefits.

What makes a "good" forecast has been the subject of many discussions throughout the history of forecasting (e.g. Peirce 1884; Clayton 1889; Nichols 1890; Mascart 1922; Winkler and Murphy 1968; Murphy 1993, 1996 and papers therein). In an essay designed to address this question, Murphy (1993) provided three distinct measures of forecast "goodness": consistency, quality, and value. Consistency, sometimes referred to as type 1 "goodness", is a measure of correspondence between the actual forecast and the forecaster's judgment of what will occur. Quality, sometimes referred to as type 2 "goodness", is a measure of correspondence between the forecast and the observations. Value, known as type 3 "goodness", is a measure of the benefit a forecast provides to users of the forecast.

A natural consequence of maximizing consistency is the need for probabilistic forecasts. This is the result of the uncertainty in a forecaster's judgment that must be conveyed in the resulting forecast. When forecasters produce probabilistic forecasts that accurately

[1] Murphy (1991) defined a rare meteorological event as one that occurs on less than five percent of forecasting occasions.

depict their uncertainty, the best expected quality, as measured by a strictly proper scoring rules (Winkler and Murphy 1968), is also achieved, cementing the need for forecasts to have a probabilistic component.

Prior to the twentieth century, weather forecasts were largely the result of rules of thumb, with little understanding for the physical mechanisms governing the atmosphere. However, several attempts at utilizing probabilities (and odds) in weather forecasts date back to the late eighteenth century when J. Dalton included measures of uncertainty in his forecasts (Dalton 1793; Murphy 1998). Although the exact forecast methods of Dalton are not known, Dalton's forecasts include statements such as "the probability of rain was much smaller than at other times" and "the probability of a fair day to that of a wet one is as ten to one," (Murphy 1998). Nearly a century later, in 1871, under the direction of Professor Cleveland Abbe, the first forecasts and warnings issued by the United States Signal Service were actually labeled probabilities. (Whitnah 1961; Murphy 1998). In fact, Professor Abbe promoted the use of "probabilities" so much he earned the nickname "Old Probabilities" (Scott 1873/1971; Murphy 1998).

Although probability forecasting of some variety was performed prior to the beginning of the twentieth century, the start of probability forecasting is often attributed to the work of W. E. Cooke (Murphy 1998). Cooke (1906) eloquently summarized the shortcomings of deterministic forecasting, and championed the inclusion of uncertainty information, by stating:

> "(a)ll those whose duty it is to issue regular daily forecasts know that there are times when they feel very confident and other times when they are doubtful as to the coming weather. It seems to me that the condition of confidence or otherwise forms a very important part of the prediction, and ought to find expression. It is not fair to the forecaster that equal weight should be assigned to all his predictions and the usual method tends to retard that public confidence which all practical meteorologists desire to foster. It is more scientific

> and honest to be allowed occasionally to say 'I feel very doubtful about the weather for tomorrow'... and it must be... useful to the public if one is allowed occasionally to say 'It is practically certain that the weather will be so-and-so tomorrow"' (p. 23; Murphy 1998).

Cooke (1906) went on to propose a forecasting paradigm in which a forecaster assigned a set of uncertainty weights to each forecast. The uncertainty weights were designed to express various levels of confidence, and it was shown that forecasters were were able to assign weights that appropriately reflected their uncertainty. Although these forecasts conveyed uncertainty information, they were not fully probabilistic in nature. It wasn't until Hallenbeck (1920) that numerical probabilistic forecasts were recorded in the literature.

Around the same time as the work done by Hallenbeck, a series of papers by Anders Ångström poignantly made the case for probabilistic rather than binary or categorical warnings (Ångström 1919, 1922; Liljas and Murphy 1994; Murphy 1998). The crux of Ångström's arguments revolved around the difficulties forecasters face when they are constrained to issue a warning in binary terms. This difficulty, argued Ångström, arises from the fact that forecasters' knowledge of users' cost/loss ratios is generally insufficient to determine whether to "issue a warning" or "not issue a warning" in the context of producing the best possible forecast for users. Ångström further noted that "(w)hat makes the matter still more complicated is the fact that... [the loss function]... has very different values in different cases," (Ångström 1922; Murphy 1998). Ångström ultimately concluded: "The most appropriate system seems therefore to be to leave to the clients concerned by the warning to form an idea of the ratio [cost-loss ratio]... and to issue the warnings in such a form that the larger or smaller probability of the event gets clear from the formulation. The client may then himself consider if it is worthwhile to make arrangements of protection or to disregard a given warning," (Ångström 1922; Murphy 1998). In these two short papers, Ångström succinctly argued the case for probability forecasts. Unfortunately, advancements in probability forecasting lay dormant until mid-century.

In addition to probability forecasting, the start of the twentieth century marked the conception of numerical weather prediction. In 1901 Professor Abbe proposed that the atmosphere was governed by a set of dynamic and thermodynamic equations, and that these equations could be used to predict future states of the atmosphere (Abbe 1901). Bjerknes (1904) followed with the recognition that in order to produce forecasts of subsequent states of the atmosphere based on the governing equations:

1. One has to know with sufficient accuracy the state of the atmosphere at a given time; and

2. One has to know with sufficient accuracy the laws according to which one state of the atmosphere develops from another.

Bjerknes recognized that the creation of a sufficiently accurate analysis of the atmosphere is a necessary condition for reliable forecasts of future atmospheric states based on the integration of the governing equations.

Lewis Richardson took up the challenge of weather prediction in the 1910s and went on to produce the first numerical prediction (i.e. one using mathematical equations; Lynch 2008). The forecast took about 6 weeks to produce by hand and predicted the change in surface pressure over the course of 6 hours. Although the forecast was a spectacular failure — predicting a 145 mb surface pressure fall when the barometric pressure remained nearly constant — Richardson considered it "a fairly correct deduction from a somewhat unnatural initial distribution," (Lynch 2008). Ultimately, Richardson's forecast was doomed from the start, falling victim to Bjerknes' first requirement for successful numerical forecasts: adequate initial conditions.

In the late 1940s, Jule Charney demonstrated that larger-scale atmospheric motions could be sufficiently predicted by advecting the geostrophic vorticity with the geostrophic wind (Charney 1947). A few years later, Charney, and his team of researchers, went on to produce the first successful numerical weather forecasts produced from a computer (Char-

ney et al. 1950). Word of this success spread rapidly throughout the meteorological community, and by the mid-1950s short-term operational numerical weather forecasts were being produced in the United States and Sweden (Lewis 2005). Although operational numerical forecasts began in the mid 1950s, these forecasts were based on the simplified equations formulated by Charney (1947). As a result of the simplified equations used in numerical weather forecasts, some meteorologists began to explore statistical methods to correct the numerical output (Gleeson 1961). Based on the work of Hinkelmann (1951), operational numerical forecasts began using the primitive equations to produce numerical forecasts in the mid 1960s (Lynch 2008).

Unfortunately, raw deterministic model output (i.e. output that has not been post-processed) is not probabilistic in nature. Deterministic model output provides a precise answer at every grid point for all output fields. This is a consequence of the internal numerics of deterministic models; equations are constrained to produce a single value. However, raw deterministic model output can still be thought of in terms of probabilities. A single output value can be interpreted as having a probability of 1 that the output value will be observed, and a 0 that any other value will be observed. Unless the deterministic model is always perfect (or always wrong), the probability that the model's forecast will be correct lies somewhere between 0 and 1.

In the 1960s, Edward Lorenz published a series of papers that demonstrated the chaotic nature of the atmosphere (Lorenz 1963, 1965, 1968). Lorenz demonstrated that even small errors in the initial state of the modeled atmosphere — even those within observational error — can have large impacts on the resulting numerical forecast. These errors abound from all aspects of numerical weather prediction, ranging from numerical to observational to the model itself. Thus, perfect, deterministic, numerical weather predictions are practically impossible. Therefore, in order for a numerical weather forecast to satisfy type 2 "goodness" they, too, must include a probabilistic component.

Armed with the results of Lorenz (1963, 1965, 1968), alternatives to the determinis-

tic numerical prediction paradigm began to be offered. Epstein (1969) recognized that the atmosphere could not be completely described with a single numerical forecast due to inherent uncertainty. Epstein (1969) and Gleeson (1970) each offered ways to produce probability forecasts in which the numerical model output means and variances directly. Although these methods showed promise, the addition of uncertainty terms proved to be too computationally expensive for operational use at that time.

Leith (1974) offered the first attempt at Monte Carlo forecasting, or ensemble forecasting as it is known today. Leith suggested that instead of running one forecast from a single initialization, a collection (ensemble) of forecasts, each with a slightly different initial state, should be produced. The resulting forecasts could be used to assess the uncertainty of the forecast. One such method of assessing the uncertainty is to report the fraction of ensemble members that produce a certain outcome divided by the total number of ensemble members. This results in a probability of occurrence, derived from the collection of numerical forecasts. As was the case with Epstein (1969) and Gleeson (1970), even though this approach showed promise, it was too expensive for operational use at that time. It took nearly two decades worth of computational advancement, but operational ensemble prediction systems became a reality in the early 1990s.

In the mean time, a different approach was put forth by Glahn and Lowry (1972). Instead of adding additional terms to the numerical model, or producing multiple model forecasts each starting with slightly different initial conditions, the idea was to post-process the raw model output using statistical models. The statistical models are in essence multivariate regressions, derived from the three-dimensional numerical output, observations, and the general climatological conditions for specific locations. These model output statistics, or MOS, account for model biases as well as local effects that cannot be resolved by the native resolution of the model. Although some MOS products are generally expressed deterministically (such as the maximum and minimum 6 hr temperatures), some products are probabilistic (such as probability of precipitation). MOS is still used today by the National

Weather Service (Allen and Erickson 2001; Allen 2001; Sfanos 2001; Carroll 2005; Glahn et al. 2009).

As previously alluded, during the latter parts of the twentieth century computing power rapidly increased. This led to a debate about whether extra resources should be devoted to increasing resolution or introducing ensembles. Operational numerical weather prediction centers committed most of the additional resources to decreasing model grid spacing (e.g, McPherson 1991; Word Meteorological Organization 1992).

The choice to embrace higher resolution numerical models resulted in much needed improvements to the models' physics parameterizations. Additionally, improved model grid spacing allowed for numerical models to better represent atmospheric phenomenon with strong gradients (i.e., fronts, drylines, inversions, etc.). These improvements allowed for highly specific spatial and temporal forecasts were easily produced (Droegemeier 1990). Unfortunately, the details in high resolution forecasts are often the least skillful aspects. These details are best in environments in which the atmosphere is dominated by large-scale phenomenon (i.e., when quasi-geostrophy can be assumed; Antolik and Doswell III 1989). Thus, numerical guidance tends to be at its best when the forecasting situation is, in some sense, the easiest, and is least needed (Brooks and Doswell 1993).

Probabilistic prediction becomes more important as the resolution of models increases, thus ensemble prediction concepts continued to receive attention if not operational implementation. (e.g. Brooks et al. 1992; Brooks and Doswell 1993). As previously noted, Lorenz (1963, 1965, 1968) demonstrated that a numerical weather forecast was sensitive to the initial conditions. Unfortunately, model grid spacing is typically less than that at which we observe the atmosphere over the globe. Thus, in addition to the errors introduced by the observing instruments, additional errors can be introduced as a result of interpolating observations to grid points that do not have collocated observations. A consequence of this is that the true state of the atmosphere is rarely, if ever, known.

Numerical weather prediction, and in particular, high-resolution models of thunder-

storms, has been shown to be very sensitive to small changes in the initial conditions (Lorenz 1963, 1965, 1968; Brooks et al. 1992; Brooks 1992). As a result, the inherent uncertainty in the observations could result in major errors in a forecast from models on that scale. An ensemble approach recognizes the uncertainty in the initial conditions and utilizes it to produce a collection of forecasts, rather than assuming that the initialization is correct (Brooks et al. 1992).

History has shown that a compromise of both of the increased resolution and ensemble forecasting paradigms have prevailed. Operational centers in the United States currently run a global ensemble at relatively coarse resolution, whereas a suite of limited area models are run at increasingly higher resolution as the domain size decreases. Recently, the computational power has increased to the point of allowing operational limited area model forecasts at grid spacings small enough for the cumulus parameterization scheme to be turned off. Even with the sensitivities to initial conditions previously mentioned, these convection-allowing models (CAMs) have shown improved skill, compared to parameterized-convection models, in identifying regions where rare meteorological events associated with convection (hereafter RCEs[2]) may occur (Clark et al. 2010). Furthermore, CAMs are able to do this by explicitly representing deep-convective storms and their unique attributes — not just storm environments (Kain et al. 2010).

Yet, as promising as CAM forecasts are, quantifying the uncertainty associated with explicit numerical prediction of RCEs is particularly challenging (Sobash et al. 2011). Of course, ensembles are powerful tools for quantifying uncertainty, but when convection-allowing ensemble prediction systems are used to provide guidance for forecasting storm-attributes, they are subject to the same fundamental limitation that handicaps single-member CAMS forecast systems: Too little is known about the performance characteristics of CAMs in predicting RCEs explicitly.

There are three main reasons for this deficiency. First, routine, explicit, contiguous

[2]Rare Convective Event

or near-contiguous United States (CONUS or near-CONUS) scale forecasts of RCEs have been available for only 6-7 years in the United States, so there is still much to learn about which phenomena can be skillfully predicted with convection-allowing models (Kain et al. 2008, 2010). Second, most real-time forecasting efforts with convection-allowing models have been short-term initiatives, focusing on specific tasks (e.g., Done et al. 2004; Weisman et al. 2008). Third, there is a limited database of forecasts for RCEs, making robust statistical techniques difficult (e.g., Hamill and Whitaker 2006). In short, there is a limited track record in the use of CAMs as guidance for prediction of RCEs.

A strategy for calibrating, or quantifying the uncertainty of, forecasts of RCEs based on the idea of generating probabilistic forecasts from a single underlying deterministic model follows. This technique uses a conceptual approach similar to that described by Theis et al. (2005) and refined by Sobash et al. (2011). As in these two studies, this strategy differs from other methods for both deterministic models (e.g., Glahn and Lowry 1972) and ensemble modeling systems (e.g., Hamill and Colucci 1998; Raftery et al. 2005; Clark et al. 2009; Glahn et al. 2009) by including a neighborhood around each model grid point as a fundamental component of the calibration process.

This strategy is rooted in the fundamental concepts of Kernel Density Estimation (KDE), which can be used to retrieve spatial probability distributions from point observations, or, in this case, forecasts. In other words, if a model forecasts an event at point A, KDE can be utilized to gain insight into the probability that the event might occur at a nearby point. This is achieved by utilizing a statistical distribution to redistribute the total probability (100%) from point A over multiple (typically nearby) grid points. The result is a probability forecast, the character of which is determined by one's choice of statistical distribution and the number of grid points over which the distribution is applied. The resulting smoothing effect is similar to that obtained with ensemble output by Wilks (2002), but initial calibration efforts herein focus on output from a single deterministic model. Sobash et al. (2011) demonstrated with a two-dimensional, isotropic Gaussian function that cal-

ibration of the probability forecasts derived using this technique is most easily done by changing the number of grid points over which non-zero probabilities are distributed. Here, however, an objective calibration method based on past model performance is presented.

The layout of this dissertation is as follows: The method is presented in Chapter 2, followed by application of the method to a deterministic model in Chapter 3. Extension to ensembles is presented in Chapter 4 and an overall discussion concludes the dissertation in Chapter 5. An explanation of the data used for both the deterministic and ensemble methods can be found in their respective chapters. 

Chapter 2

Proposed Method

In theory, a perfect deterministic numerical forecast would consist of a perfect (in amplitude and phase) initialization and a perfect integration forward in time resulting in the correct forecast of what was observed. Unfortunately, the current state of our modeling systems is such that perfect forecasts are impossible. Numerical forecasts begin with errors in the initialization, use imperfect approximations of physical processes, and utilize discretized approximations of the continuous atmosphere, all of which result in errors in the final forecast.

This inherent uncertainty in numerical forecasts has led to the recognition that numerical forecasts should have a probabilistic component (e.g. Glahn and Lowry 1972; Murphy 1993; Glahn et al. 2009). For many years probabilistic guidance has been produced from deterministic forecasts in the United States using model output statistics (commonly referred to as MOS), a logistical regression process in which the likelihood of a specified outcome at a given location is estimated on the basis of past model performance (Glahn and Lowry 1972). Unfortunately, statistical post-processing of model output works best when modeling systems allow for the creation of a forecast sample that adequately represents the larger forecast population. In the case of operational numerical models this implies modeling systems must remain consistent for long periods of time. Calibration becomes more challenging with modeling systems that have been recently modified, which limits the forecast sample. Additional challenges arise when dealing with rare events, which require a long forecast sample to adequately represent the forecast population of rare events.

The statistical post-processing method proposed here goes beyond Theis et al. (2005) and Sobash et al. (2011) by employing a compositing and fitting technique for objective calibration of forecast probabilities. The idea behind the proposed method is to objectively determine where observations of a given event occur relative to forecasts of that event. This information is then used to fit an analytic function to the two-dimensional frequency histogram. Several analytic functions might be good candidates for this purpose; a two-dimensional Gaussian function is used here, fitted using methods similar to Lakshmanan and Kain (2010). In theory a normalized version of the raw two-dimensional frequency histogram, could be used, but, there is no guarantee that this would produce continuous forecast probabilities. After fitting an analytic function to the two-dimensional histogram, this analytic function is used as the kernel to transform a forecast of "yes"/"no" occurrence into a continuous probabilistic forecast using methods similar to Kernel Density Estimation (Silverman 1986).

2.1 2-D Compositing

The compositing employed here is straight forward. First, forecasts are mapped onto the observational grid. Next the raw forecasts and observations are converted into binary grids of 1s (event occurred) and 0s (event did not occur). Next, all grid points within a specified radius of a forecast event are examined for observed events. If an observed event occurs within the specified radius of a forecast, the position of the observation relative to the forecast is recorded. Lastly, all of these relative positions are mapped onto a common grid and accumulated over all forecast periods during a specified training period. The result is a two-dimensional frequency histogram of where observations occurred relative to forecasts during the entire training period.

2.2 Fitting

Once the two-dimensional frequency histogram is determined, the next step involves fitting an analytic function to the histogram. As previously mentioned, any analytic function would work, as long it accurately modeled the two-dimensional frequency histogram. A Gaussian function is demonstrated here.

To fit a Gaussian function, two parameters must be known: location and variance. Typically, the location of a Gaussian function is given by the mean of the distribution. In the case of a two-dimensional histogram this is given by the weighted center, or centroid, of the distribution. Details on how to compute centroid are provided in subsection 2.2.1. In the case of a two-dimensional Gaussian, the variance must be computed in both the X- and Y-dimensions. Details on this are provided in subsection 2.2.2.

2.2.1 Finding the 2-D Histogram's Centroid

After creation of the two-dimensional frequency histogram, the next step involves determining the centroid of two-dimensional frequency histogram, and, in turn, determining the mean displacement vector. The centroid's location, represented by μ_x and μ_y can be computed using

$$\mu_x = \frac{1}{\mathbf{H}} \sum_{\substack{j=-m \\ i=-n}}^{n,m} H_{ij} \cdot i, \tag{2.1}$$

$$\mu_y = \frac{1}{\mathbf{H}} \sum_{\substack{j=-m \\ i=-n}}^{n,m} H_{ij} \cdot j, \tag{2.2}$$

where H is the observed two-dimensional histogram, n is half the length of the x-dimension of H, m is half the length of the y-dimension, and

$$\mathbf{H} = \sum_{\substack{j=-m \\ i=-n}}^{n,m} H_{ij}, \tag{2.3}$$

given that the forecast is found at $(0,0)$. Thus, the observation mean displacement vector is simply (μ_x, μ_y).

2.2.2 Finding the 2-D Histogram's Standard Deviation

Once the centroid of the two-dimensional histogram is known, the next step to fitting a two-dimensional Gaussian function is to compute the variance in both dimensions. The two-dimensional Gaussian function is represented mathematically as:

$$G(x, y, \sigma_x, \sigma_y) = \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}\right)}, \tag{2.4}$$

where σ_x is the standard deviation (or, bandwidth, when applied in a KDE framework) in the x-direction and σ_y is the standard deviation in the y-direction. When $\sigma_x \neq \sigma_y$, the Gaussian function is said to be anisotropic. The function's peak probability density occurs at the center of the function, and decreases monotonically outward such that

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1, \tag{2.5}$$

and

$$\frac{a}{b} = \frac{\sigma_x}{\sigma_y}, \tag{2.6}$$

where x is the x-component of the distance from the center, y s the y-component of the distance from center, a is the major axis, and b is the minor axis, is satisfied. The result is a set of concentric ellipses that monotonically decrease outward. In the limiting case of $\sigma_x = \sigma_y$, the Gaussian function is said to be isotropic. In this case, the function's peak

probability density still occurs at the center of the function and decreases monotonically outward; however, equation (2.5) becomes

$$\frac{x^2+y^2}{r} = 1, \tag{2.7}$$

where $r = \sqrt{a^2+b^2}$. In this case, concentric circles are produced rather than ellipses.

In Sobash et al. (2011) attempts to produce reliable forecasts centered on subjectively changing σ_x for a two-dimensional isotropic Gaussian function until the value producing the most reliable forecasts was found[1]. Here, the proposed calibration method uses a two-dimensional anisotropic Gaussian function, with σ_x and σ_y objectively derived from the model's historical error characteristics. This is done by using

$$\sigma_x = \frac{1}{\mathbf{H}} \sum_{\substack{j=-m \\ i=-n}}^{n,m} H_{ij}(i-\mu_x)^2, \tag{2.8}$$

$$\sigma_y = \frac{1}{\mathbf{H}} \sum_{\substack{j=-m \\ i=-n}}^{n,m} H_{ij}(j-\mu_y)^2, \tag{2.9}$$

which can then be used plugged into equation (2.4).

Unfortunately, the two-dimensional Gaussian function represented by equation (2.4) forces σ_x and σ_y to lie along the abscissa and ordinate, respectively. In order to model a two-dimensional distribution where the major and minor axes are rotated off the abscissa and ordinate, a rotation matrix,

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}, \tag{2.10}$$

where θ is the rotation angle (in radians) of the major axis from the abscissa in the counter-clockwise direction, must be applied to the two-dimensional Gaussian given in

[1] In a 2D isotropic Gaussian, $\sigma_x = \sigma_y$, thus equation (2.4) is reduced to a form only depending on σ_x

equation (2.4). The rotated, two-dimensional Gaussian equation can be represented mathematically by

$$G'(x', y', \sigma_x, \sigma_y) = \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}\left(\frac{x'^2}{\sigma_x^2} + \frac{y'^2}{\sigma_y^2}\right)}, \tag{2.11}$$

where x' and y' are determined by equation (2.10).

The rotation angle can be computed from the covariance matrix of the observed frequency distribution,

$$\Sigma_{xy} = \begin{bmatrix} \sigma_x & \sigma_{xy} \\ \sigma_{xy} & \sigma_y \end{bmatrix} \tag{2.12}$$

where σ_{xy} is the covariance of x and y, and is given by

$$\sigma_{xy} = \frac{1}{\mathbf{H}} \sum_{\substack{j=-m \\ i=-n}}^{n,m} H_{ij}(i - \mu_x)(j - \mu_y), \tag{2.13}$$

by taking the arctan $\left(\frac{y''}{x''}\right)$, where x'' and y'' are the x- and y-components of the eigenvector of Σ_{xy} ~~containing~~ the maximum eigenvalue (Lakshmanan and Kain 2010).

2.3 Putting it All Together

Once μ_x, μ_y, σ_x, σ_y, and θ are known, kernel density estimation can be carried out on any forecast. This is done by creating a binary grid from the model's forecast, similar to the ones created in the compositing stage. Each "yes" grid point (1) is then shifted by the mean displacement vector, μ_x, μ_y and then smoothed by equation (2.11), using the objectively derived σ_x, σ_y, and θ out to a distance of $5\sigma_x$ in the x-direction and $5\sigma_y$ in the y-direction[2]. This generates probabilities at the nearby grid points that the corresponding observation

[2] $5\sigma_x$ and $5\sigma_y$, were chosen because this distance accounts for greater than 99.999% of the fitted Gaussian distribution. In practice, any multiple of σ_x and σ_y would work, with larger multiples accounting for more of the fitted Gaussian's tails.

will actually occur there, rather than solely at the location of the forecast event. The final forecast probabilities then become the linear combination of all the forecast probabilities created by applying the fitted Gaussian to each forecast event.

Chapter 3

Deterministic

3.1 Deterministic Data

To develop a statistical post-processing method for calibration of numerical model forecasts, forecasts and corresponding observations must be readily available. Unfortunately, observational datasets of many RCEs (e.g., tornadoes, wind damage, hail swaths, etc) are riddled with inaccuracies and deficiencies (Doswell and Burgess 1988; Weiss et al. 2002; Trapp et al. 2006; Ortega et al. 2009) that prevent their use in rigorous statistical post-processing. Precipitation, however, is one of the best documented observational fields in meteorology, and it is available as direct output from numerical models. Thus, to develop a method of calibration for RCEs, it was decided to focus on precipitation. Here, accumulations greater than or equal to 25.4 mm in 6 hr are considered rare events as this threshold is met on fewer than 0.36% of all grid points[1] (Figure 3.1). Subsequently, the proposed method was applied to forecasts of accumulations greater than or equal to 12.7 mm in 6 hr to test the method against a less rare event[2] (Figure 3.2).

Model forecasts and observations of precipitation were obtained for the 48-month period 01 April 2007 through 31 March 2011 and subdivided into two categories: training data and verification data. Forecasts and observations during the time period 01 April 2007 through 31 March 2010 (36 months) were used in the training dataset, with the remaining

[1] If one considers a forecast at a single grid point as a single forecasting occasion, then 25.4 mm in 6 hr meets the definition of a rare event posed by Murphy (1991).

[2] Forecasts of at least 12.7 mm in 6 hr occurred on approximately 1.3% of all grid points.

12 months used for testing and verification of the proposed method.

3.1.1 Deterministic Model Forecasts

Model forecasts were taken from the 4 km grid-length Weather Research and Forecasting – Advanced Research WRF (WRF-ARW) model configuration (Skamarock et al. 2008) run daily at the National Oceanic and Atmospheric Administration (NOAA) National Severe Storms Laboratory (NSSL). The NSSL produces numerical weather prediction forecasts from the WRF-ARW model as part of an ongoing collaborative effort with the NOAA Storm Prediction Center (SPC). Model forecasts are produced daily, integrated out to 36 hours, using 00 UTC initial and lateral boundary conditions from the operational North American Mesoscale model (Rogers et al. 2009), over a CONUS domain. Information on the configuration is provided in Kain et al. (2010).[3]

3.1.2 Observations

Observations were taken from the NOAA National Centers for Environmental Prediction (NCEP) Stage IV national quantitative precipitation estimate analysis. The stage IV analyses are based on the multi-sensor hourly/6-hourly 'Stage III' analyses (on local 4.7 km polar-stereographic grids) produced by the 12 River Forecast Centers in the CONUS. NCEP mosaics the Stage III into a national product (the Stage IV analyses) available in hourly, 6-hourly, and 24-hourly (accumulated from the 6-hourly) intervals. Lin and Mitchell (2005) describe further details of these analyses[4].

[3]Images of output from the WRF forecasts generated at the NSSL, hereafter NSSL-WRF, can be found at `http://www.nssl.noaa.gov/wrf`

[4]Archives of the Stage IV dataset can be found at `http://data/eol.ucar.edu/codiac/dss/id=21.093`.

3.1.3 Processing

Diagnostic analyses were conducted on the Stage IV grid, requiring interpolation of the NSSL WRF-ARW[5] output. The program *copygb*[6] was used for the interpolation and domain-wide total liquid volume was conserved. Six-hour accumulation periods were used for both model forecasts and the Stage IV analyses, with model forecasts coming from the 12-36 hr forecasts ending at 18, 00, 06, and 12 UTC. A mask was applied to both the NSSL-WRF forecasts and Stage IV analyses to limit the region studied to CONUS and near-CONUS areas east of the Rocky Mountains (Fig. 3.3).

Over the course of the three years training dataset and the one year verification dataset there were occasional 6 hr time periods where either the NSSL-WRF data or the Stage IV analyses were unavailable. If either the Stage IV analysis or the NSSL-WRF data were unavailable for a given 6 hr time period, that particular time period was removed from consideration. Additionally, some of the Stage IV analyses contain missing data over portions of the domain shown in Figure 3.3. In these situations, the missing grid points from the Stage IV analyses were masked on the corresponding NSSL-WRF grid, and the training and verification of the proposed method was carried out on the remaining grid points.

3.2 Deterministic Results

3.2.1 25.4 mm Threshold

Before the proposed method can be applied to a numerical forecast, the two-dimensional histogram of observations relative to forecasts must be constructed. A distance of 400 km was used as the search radius over which to look for observed accumulations of at least 25.4 mm in 6 hr relative to forecast accumulations of the same. The resulting two-dimensional composite is shown in Figure 3.4.

[5]Hereafter referred to as NSSL-WRF.

[6]Available at `http://www.cpc.ncep.noaa.gov/products/wesley/copygb.html`.

It is readily apparent from examining Figure 3.4 that the centroid of the two-dimensional frequency distribution is observed to the north-northeast of the representative forecast grid point and the distribution has an elliptical shape. When the anisotropic Gaussian function is fit to this distribution, the resulting parameters, determined by the methods discussed in chapter 2 as well as in Lakshmanan and Kain (2010), are $(\mu_x, \mu_y) = (4.7, 18.8)$ kilometers, indicating that the NSSL-WRF forecasts were, on average, approximately 4.7 km too far west and 18.8 km too far south, $\sigma_x \approx 190$ kilometers, $\sigma_y \approx 155$ kilometers, and $\theta \approx 60^\circ$ in the counter-clockwise direction, revealing the anisotropy of the distribution. To some extent, the shape and anisotropy are closely related to the mean shape and orientation of individual precipitation objects, as revealed by comparing the average size-weighted orientation of the precipitation objects, determined by the Baldwin object identification algorithm (Baldwin et al. 2005), to the orientation angle of the fitted distribution (not shown). In particular, a southwest-to-northeast orientation of heavy precipitation is often observed.

Using this fitted two-dimensional distribution, probabilistic forecasts for each 6 hr time period from 01 April 2010 — 31 March 2011 were generated as described in Section 2.3. Four sample forecasts, all of differing lead times, are shown in Figures 3.5 and 3.6 and are now discussed. However, one must be cautious about assessing the skill of a probabilistic forecasting system on the basis of individual events.

Figure 3.5a, 3.5c, and 3.5e depict observations and model forecasts of precipitation for the 6 hours ending 18 UTC 02 May 2010 (a 12-18 hr forecast). During this 6 hr period, heavy-rain fell over an elongated area stretching from central Mississippi north-northeastward into southeastern Ohio and western West Virginia, with an area exceeding 200 mm in north-central Tennessee (Figure 3.5a). South and east of this axis of heaviest rainfall, areas in eastern Mississippi had precipitation totals around the 25.4 mm threshold. The NSSL-WRF forecast of this event was generally good, cluing forecasters in on the general area of concern. However, the NSSL-WRF forecast had three distinct areas of heavy rain compared to the single large band that was observed: one northwest of the ob-

served axis of heavy rain, one southeast, and one along the northeastern most observed area exceeding 25.4 mm (Figure 3.5c). Applying the proposed probabilistic method resulted in the area of highest probabilities of reaching or exceeding 25.4 mm (between 25 and 30%) occurring very near the area of maximum rainfall (Figure 3.5e). Additionally, the axis of highest probabilities extending northeast of the maximum probabilities aligned very well with the observed area equal to or exceeding 25.4 mm. The axis of highest probabilities also extends to the south and southwest of the maximum forecast probabilities, capturing the southwestward extent of the observed heavy rain, at the same time highlighting areas in eastern Mississippi (Figure 3.5e).

Figures 3.5b, 3.5d, and 3.5f depict observations and model forecasts of precipitation for the 6 hours ending 00 UTC 27 September 2010 (an 18-24 hr forecast). Observations depict a large area of precipitation greater than or equal to 25.4 mm stretching from southeastern Alabama northeastward into far northwestern South Carolina with scattered areas reaching this threshold across southern Mississippi and eastern North and South Carolina (Figure 3.5b). The NSSL-WRF forecast of this event depicted two areas exceeding 25.4 mm of precipitation, essentially capturing both observed areas (cf. Figures 3.5b and 3.5d). The corridor of observations greater than 25.4 mm are generally contained within 5-10% probabilities (Figure 3.5f). In this case, much of the area covered by the highest probabilities of 15-20% did not receive heavy rainfall during this period.

A 24-30 hr forecast and observations of precipitation for the 6 hours ending 06 UTC 06 June 2010 are presented in Figures 3.6a, 3.6c, and 3.6e. Observations depict two areas over Michigan that reach the 25.4 mm threshold. The first extends from the southeastern portion of Lake Michigan eastward to the western portions of Lake Erie. The second area extends from the northern portion of Lake Michigan eastward to the western portions of Lake Huron. A third area reaching the 25.4 mm threshold is found across Illinois and into Indiana (Figure 3.6a). Although slightly farther west, the NSSL-WRF deterministic forecast does a reasonable job depicting the general location of the heaviest precipitation across

Illinois and southern Michigan. However, it under-predicts the heavy precipitation across northern Michigan (Figure 3.6c). The probabilistic forecast derived from the NSSL-WRF captures most, if not all, observed areas that reached the 25.4 mm threshold with a probability of at least 5% – including the area across northern Michigan that was not explicitly forecast to exceed 25.4 mm by the deterministic forecast. Furthermore, the highest probabilities are located in southwestern Michigan (30-35%), conjoined with the western portion of the southern Michigan heavy rain axis (Figure 3.6e).

A 30-36 hr forecast and observations of precipitation for the 6 hours ending 12 UTC 30 September 2010 are presented in Figures 3.6b, 3.6d, and 3.6f. Observations depict a large area exceeding the 25.4 mm threshold extending from eastern South Carolina, northward into far southeastern New York (Figure 3.6b). Additionally, a small region of precipitation reaching the 25.4 mm threshold is found across northeastern Georgia. The NSSL-WRF deterministic forecast is slightly narrower and farther east with its forecast, misplacing the axis of heaviest precipitation across North Carolina and Virginia (Figure 3.6d). However, the NSSL-WRF generated probabilities encompass the area exceeding the 25.4 mm threshold, with the maximum probabilities of 40-45% near Washington D.C. (Figure 3.6f). The NSSL-WRF deterministic forecast completely missed the heavy precipitation across northeastern Georgia, and this area is sufficiently far from the area to the east that it falls outside the 0.1% contour of the probabilistic forecast.

These examples are illuminating but many events are required to assess the skill of probabilistic forecast systems. A more objective verification is provided here by applying well-known verification metrics to the entire 12-months worth of forecasts generated in this manner. First, all probabilistic forecasts, which are continuous, were binned into 5% bins ranging from 0 to 100. Probabilistic forecasts of 0% and 100% were placed into their own specific bins[7]. Next, the Probability of Detection (POD) and Success Ratio (SR)[8]

[7]Technically, a forecast of 100% was never achieved.

[8]The Success Ratio is defined as 1 - False Alarm Ratio (FAR).

were computed for each of the probabilistic bins. Then, the POD and SR were plotted on a performance diagram (Roebber 2009) (Figure 3.7a). The performance diagram reveals that the best probability bin has a Critical Success Index (CSI) of around 0.1. Unfortunately, most of the probability bins had a CSI much lower than 0.1.

Although the performance diagram demonstrates relatively poor results in terms of CSI, the reliability of the forecasts is a different matter (Figure 3.7b). The resulting diagram indicates that forecasts are quite reliable over a broad range of probabilities. One caveat, however, is that the variability of the reliability increases as the number of forecast grid points decreases, especially when the total number of forecast grid points approaches or falls below 100 000 (Figure 3.7c).

3.2.2 12.7 mm Threshold

After evaluating the method at a threshold of 25.4 mm in 6 hr, the method was evaluated at a lower threshold. Similar to the two-dimensional frequency histogram at the 25.4 mm threshold, Figure 3.8 conveys that the centroid of the two-dimensional frequency distribution generated from using the 12.7 mm threshold is also observed to the north-northeast of the representative forecast grid point. As was the case with the 25.4 mm threshold, the distribution at the 12.7 mm threshold also has an elliptical shape. The fitted anisotropic Gaussian function has parameters $(\mu_x, \mu_y) = (9.4, 14.1)$ kilometers, indicating that the NSSL-WRF forecasts were, on average, approximately 9.4 km too far west and 14.4 km too far south, $\sigma_x \approx 190$ kilometers, $\sigma_y \approx 170$ kilometers, and $\theta \approx 50^\circ$ in the counter-clockwise direction. The fitted function at the 12.7 mm threshold is less anisotropic as compared to the 25.4 mm threshold. This is most likely a result stemming from using a lower precipitation threshold. The lower precipitation threshold is easier to achieve, resulting in more grid points being considered as "events". The more events results in a wider composite at the 12.7 mm threshold than at the rarer 25.4 mm threshold.

Examples of the forecasts produced by the calibration method are shown in

Figures 3.9 and 3.10. The general structure of the precipitation exceeding 12.7 mm is similar to that exceeding 25.4 mm discussed in section 3.2.1 and will not be discussed in detail here. Generally speaking, the probabilistic forecasts generated at this threshold appear to be of similar quality as those shown for the 25.4 mm threshold. The main areas of precipitation exceeding the 12.7 mm threshold appears to be contained within areas exceeding probabilities of 5%. The most notable difference is the higher probabilities generated.

As mentioned previously, even though these examples appear to produce good probabilistic forecasts, it is improper to assess the quality of a probabilistic forecast system using isolated events. Instead probabilistic forecasts need to be evaluated over many events. Following the methods outlined previously, a performance diagram and reliability diagram were constructed for the forecast results at the 12.7 mm threshold (Figure 3.11). The performance diagram is much improved over that in Figure 3.7a, as all forecast bins have improved CSI; the maximum CSI attained is closer to 0.2. The reliability diagram indicates forecasts over a much broader range of probabilities than those in Figure 3.7b. The forecasts at lower probabilities demonstrate nearly perfect reliability; forecasts at higher probabilities deviate slightly from perfect reliability, with the trend toward under forecasting. In other words, when the method produces a forecast with a high probability, the event occurs more often than a reliable forecast would. As is the case with the 25.4 mm threshold, as the number of forecast grid points decreases, particularly below 10 000 grid points, the forecast is more likely to deviate from perfect reliability (Figure 3.11c versus Figure 3.7c).

3.2.3 Discussion

As is apparent, the proposed calibration technique offers a method of objectively generating calibrated probabilistic forecasts of RCEs from a single deterministic model. This technique is successful because it objectively represents, and corrects for, the mean displacement and the spatial uncertainty associated with the underlying deterministic forecast system. Preliminary assessments suggest that this uncertainty varies systematically as a

function of numerous factors, such as forecast lead time, geographic location, meteorological season and regime, etc. Further refinements to the technique could include dependencies on these factors. For example, since cool-season precipitation forecasts tend to be more accurate than those for the warm-season, Gaussian fits to the position-error fields could vary as a function of season, with sharper, higher amplitude distributions in the cool-season and broader, lower amplitude distributions in the warm-season.

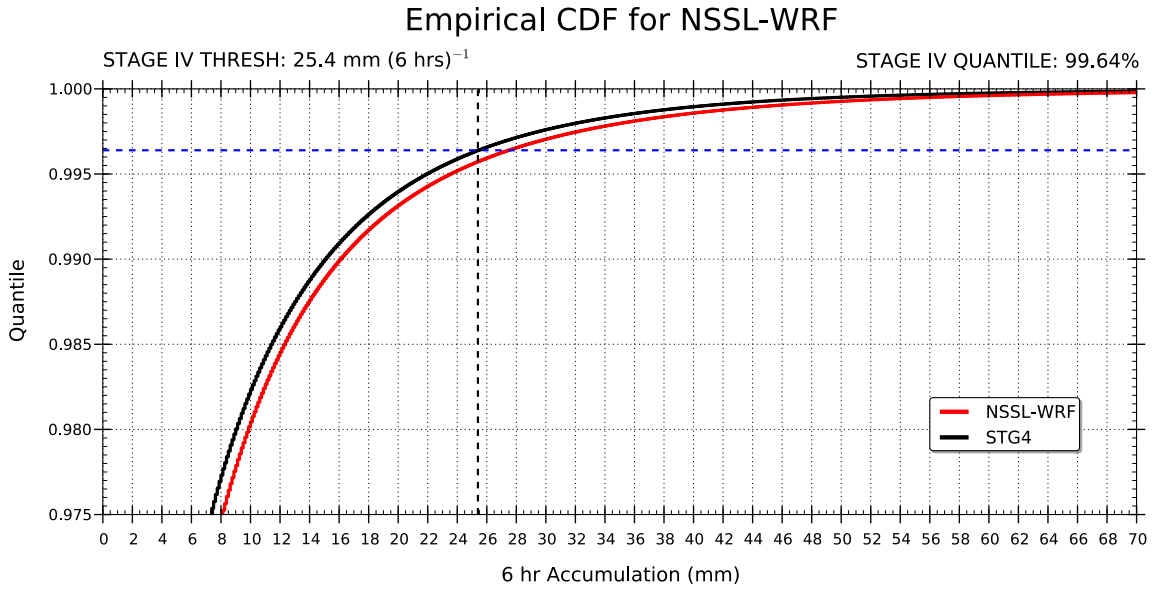

Figure 3.1: The empirical cumulative distribution function for both the Stage IV observations (black) and the NSSL-WRF forecasts (red), derived over the time period 01 April 2007 – 31 March 2010. The vertical black dashed line is the 25.4 mm threshold. The horizontal blue dashed line is the Stage IV quantile associated with the 25.4 mm threshold. Where the blue dashed line intersects the NSSL-WRF empirical cumulative distribution function is the corresponding NSSL-WRF threshold at which the ratio of points above to points below is equal to the Stage IV ratio of points above to points below the 25.4 mm threshold. This new threshold for the NSSL-WRF is 26.625 mm.

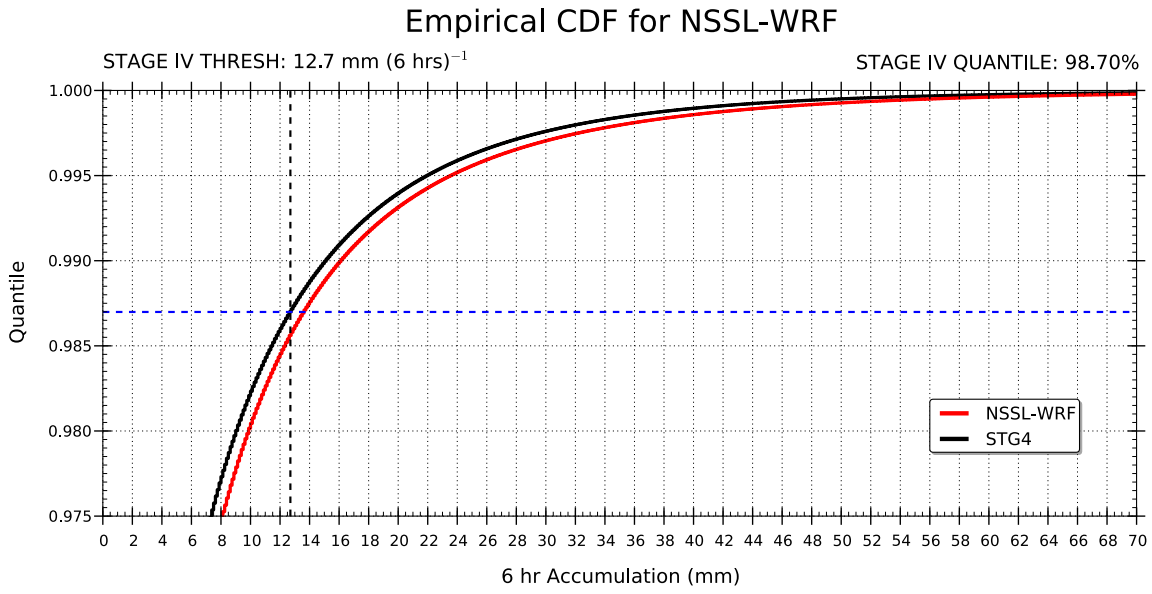

Figure 3.2: The same as in Figure 3.1, but using the 12.7 mm threshold. The new threshold for the NSSL-WRF is 13.75 mm.

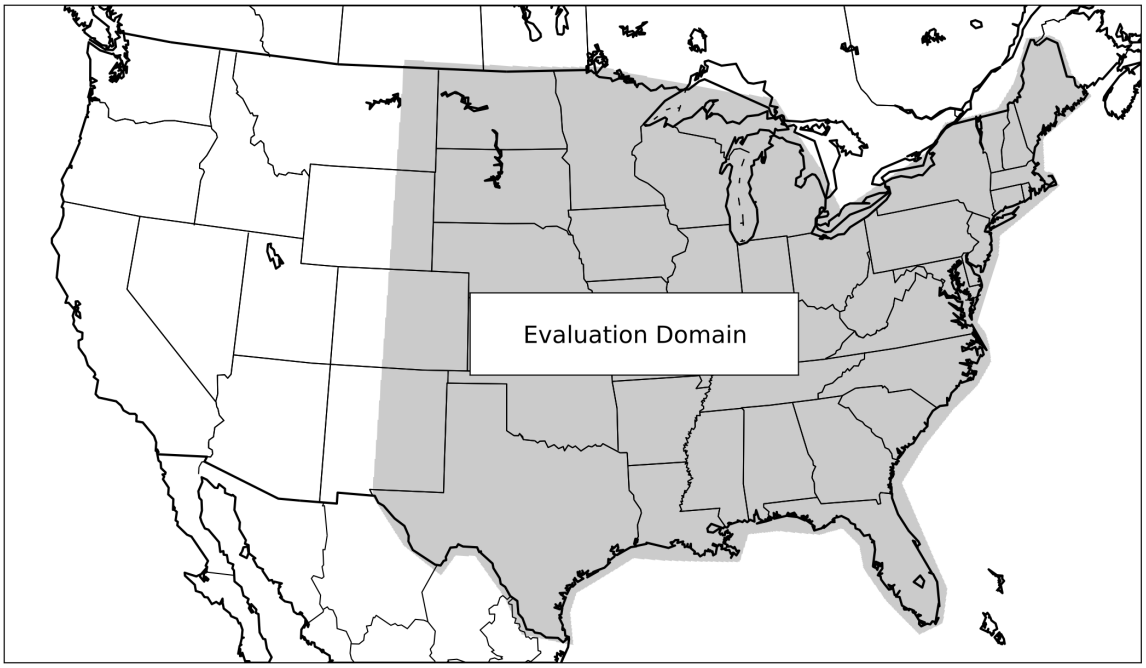

Figure 3.3: The subset of the Stage IV grid used in the analysis.

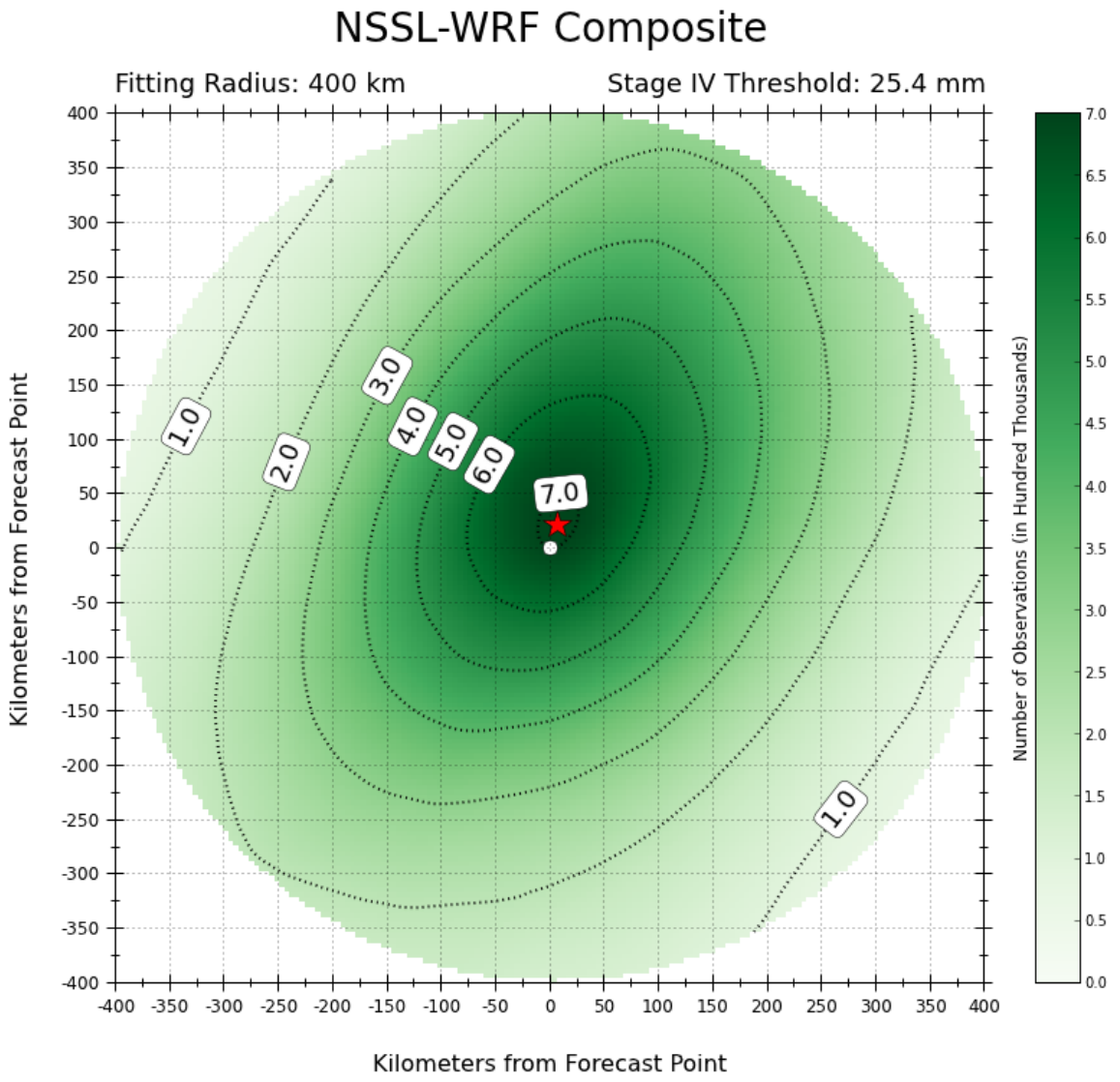

Figure 3.4: The two-dimensional frequency distribution of stage IV observations greater than or equal to 25.4 mm relative to NSSL-WRF forecasts of similar events for the training dataset (1 Apr 2007 – 31 Mar 2010). The representative NSSL-WRF forecast grid point is marked by a white dot in the middle of the domain and the stage IV observation frequency is color filled. To illustrate the displacement between forecasts and observations, the centroid of the observations is denoted by the black dot. Contour labels are given in hundred-thousands.

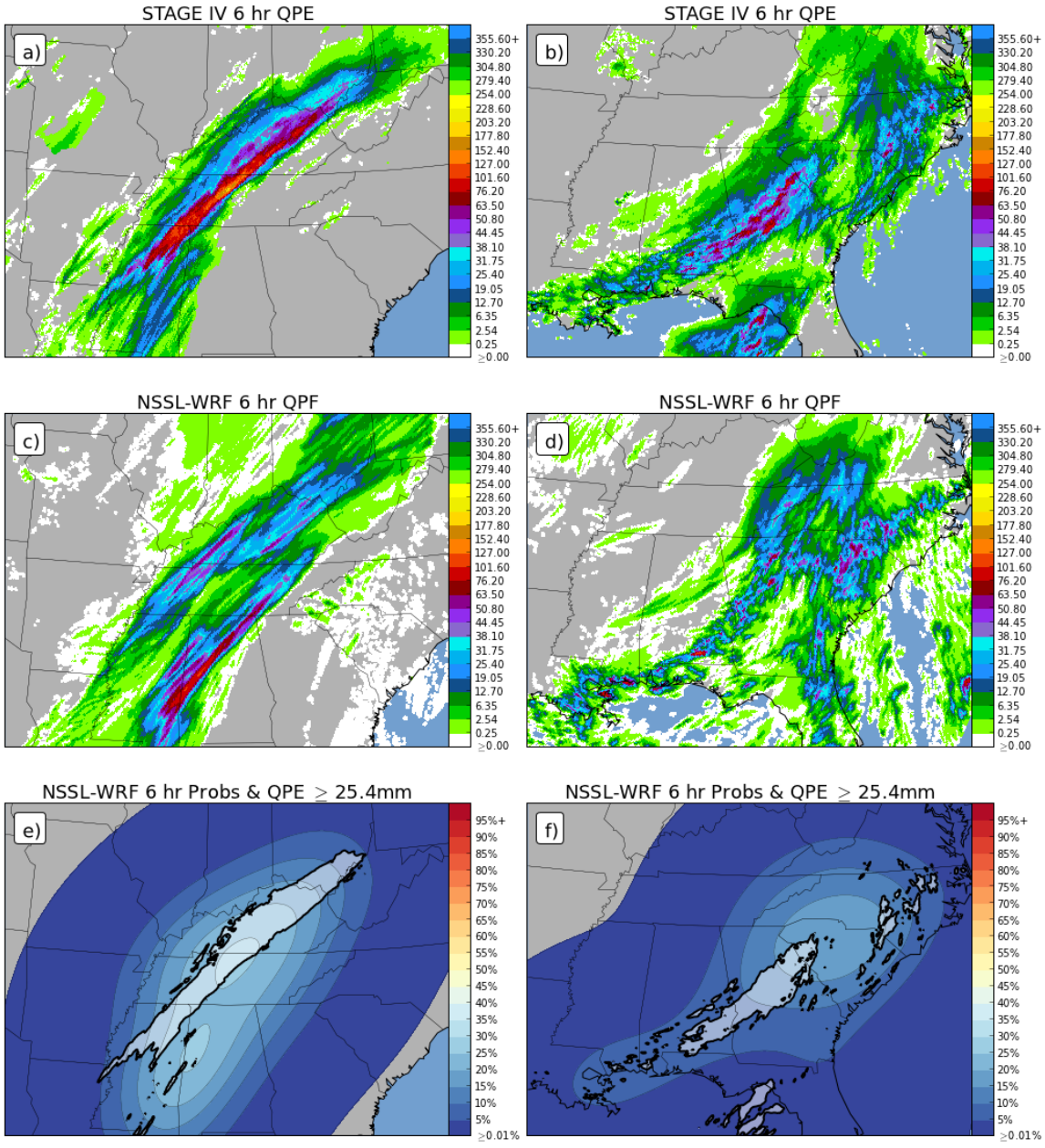

Figure 3.5: Example forecasts and observations from two separate days and differing forecast lengths. The column on the left depicts forecasts and observations for the 6-hrs ending 02 May 2010 at 18 UTC (12-18 hr forecast) whereas the column on the right depicts forecasts and observations for the 6-hrs ending 27 September 2010 at 00 UTC (18-24 hr forecast). Panels (a) and (b) denote the Stage IV 6-hr quantitative precipitation estimates (QPE), panels (c) and (d) denote the 6-hr NSSL-WRF 6-hr quantitative precipitation forecasts (QPF), and panels (e) and (f) depict the Stage IV QPE greater than 25.4 mm contoured on top of the NSSL-WRF probability of exceeding 25.4 mm in 6-hrs. The minimum shaded probability is 0.0001 (0.01%).

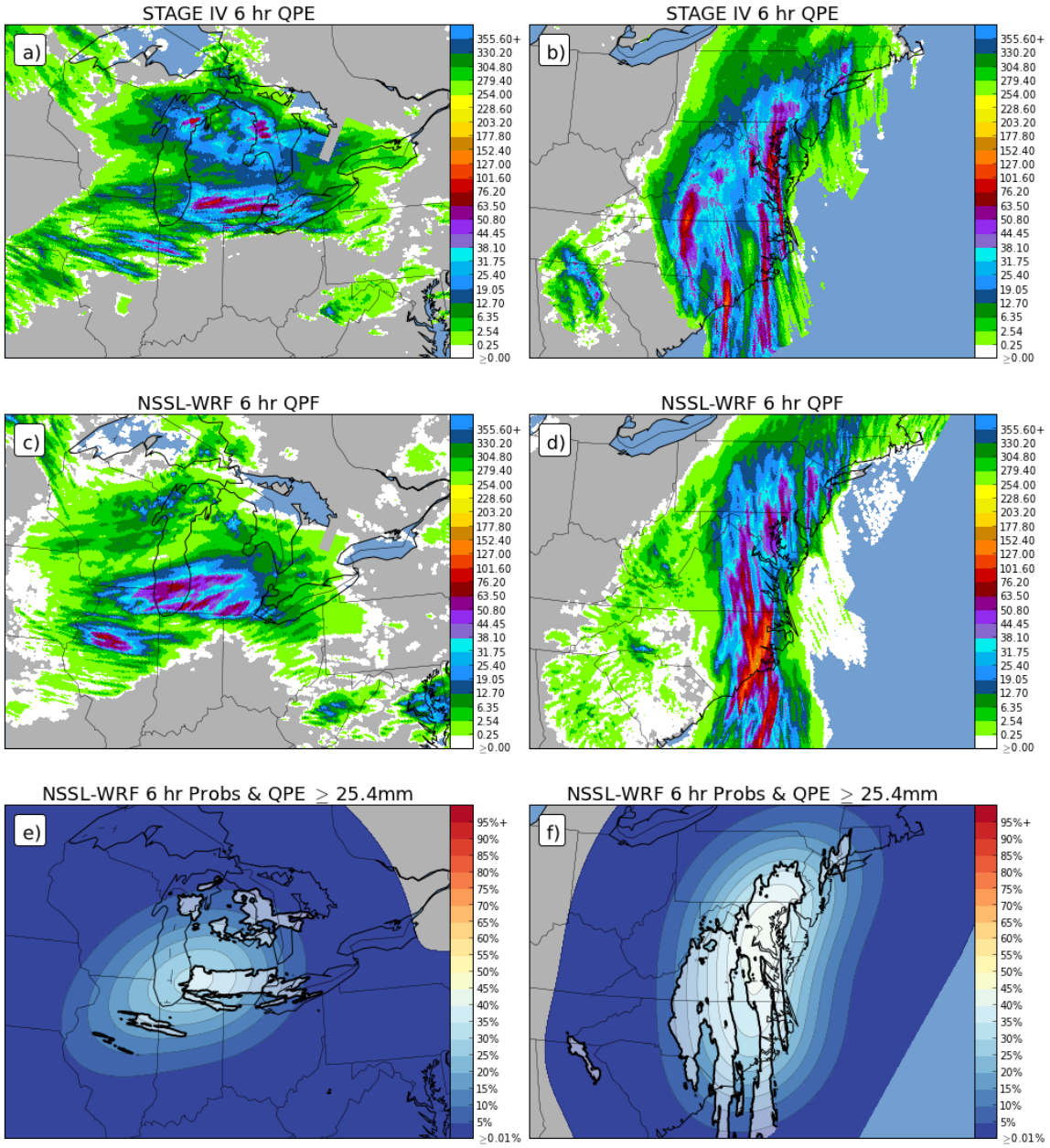

Figure 3.6: Same layout as in Figure 3.5 except the left column depicts the forecast and observations for the 6-hrs ending 06 June 2010 at 06 UTC (24-30 hr forecast) and the right column depicts the 6-hrs ending 30 September 2010 at 12 UTC (30-36 hr forecast).

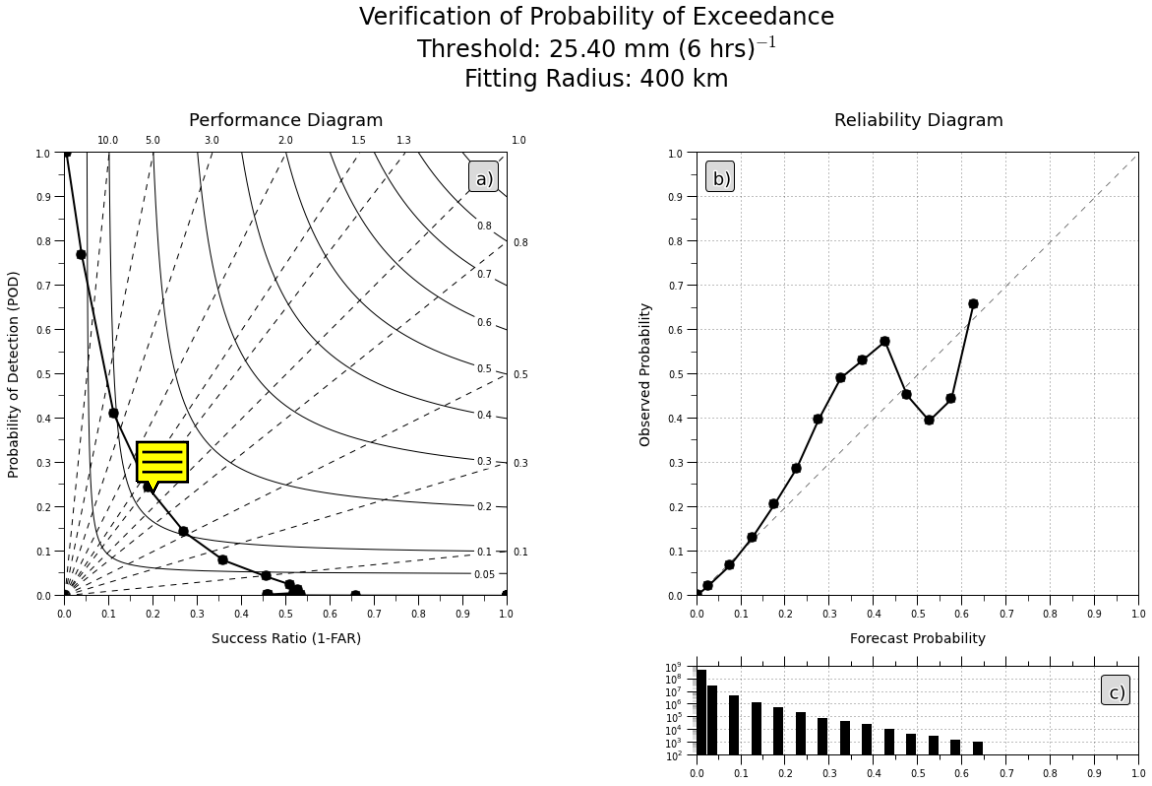

Figure 3.7: Performance Diagram (a) and reliability diagram with corresponding forecast counts (b), both computed over the 01 April 2010 to 31 March 2011 time period. The line of perfect reliability (diagonal; dashed) is also plotted on the reliability diagram. The forecast counts associated with the reliability diagram are plotted on a log-scale below the reliability diagram (c).

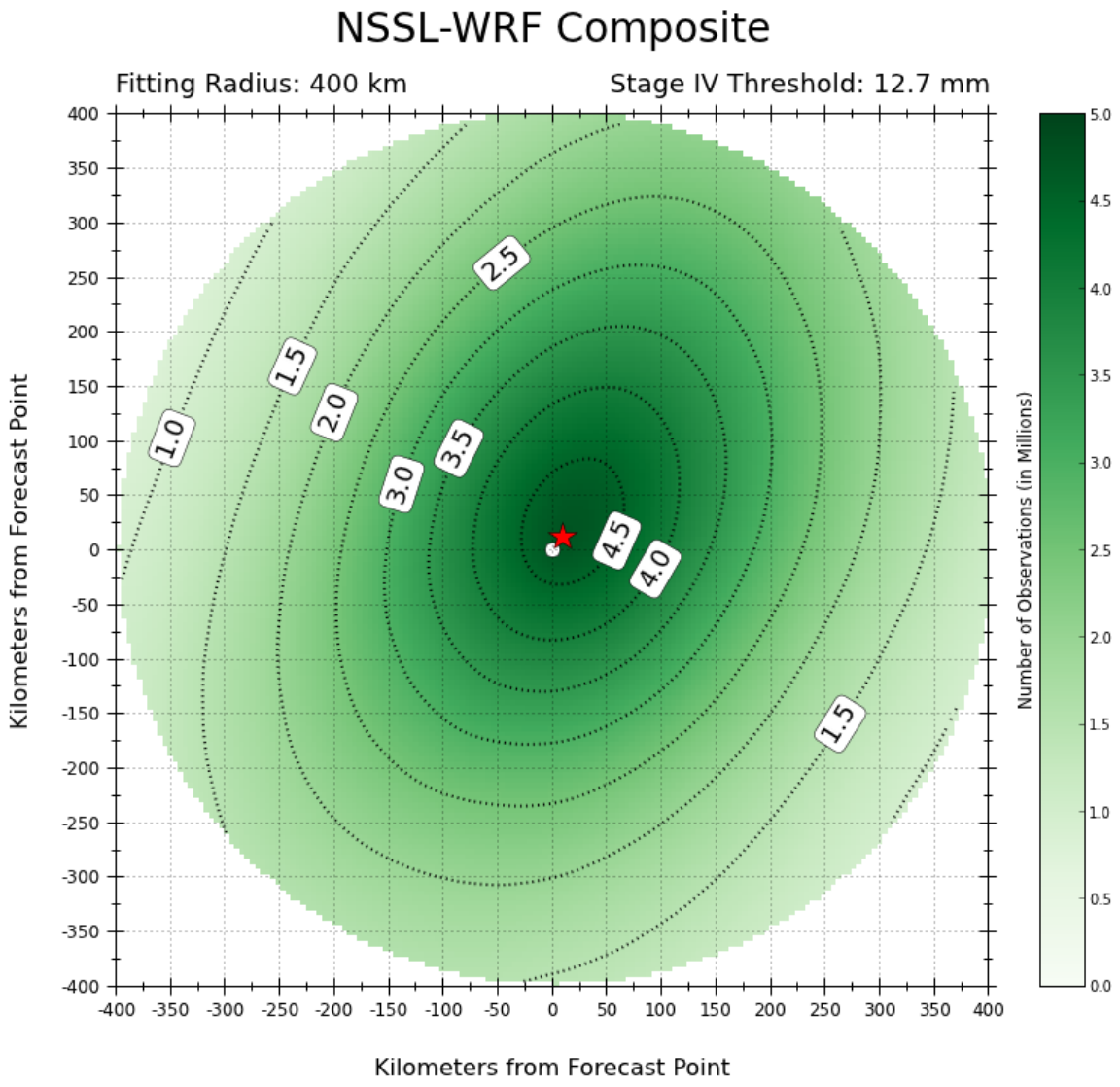

Figure 3.8: Same as in Figure 3.4 except for the 12.7 mm threshold and contour labels in millions.

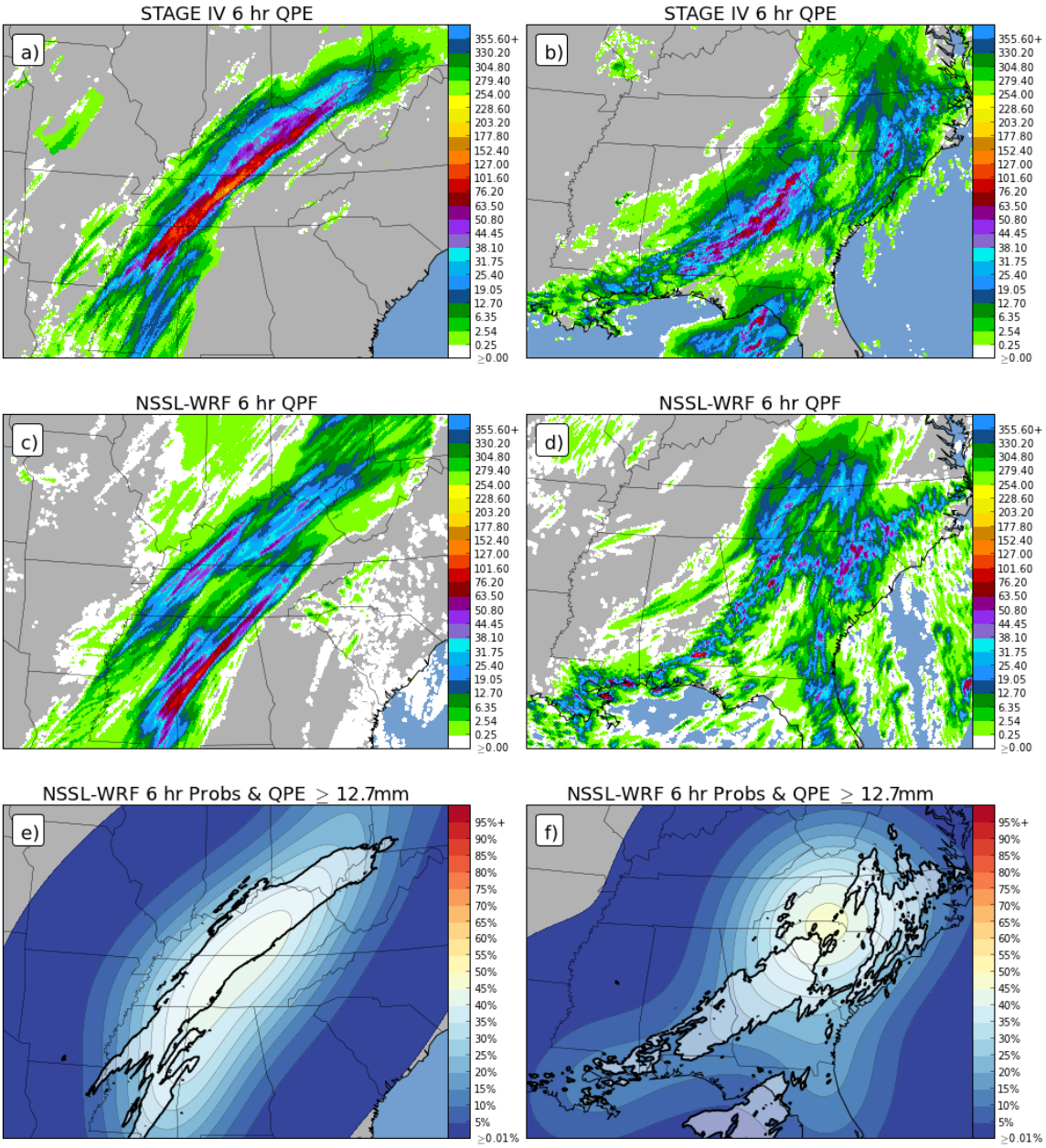

Figure 3.9: The same as in Figure 3.5 except using the 12.7 mm threshold.

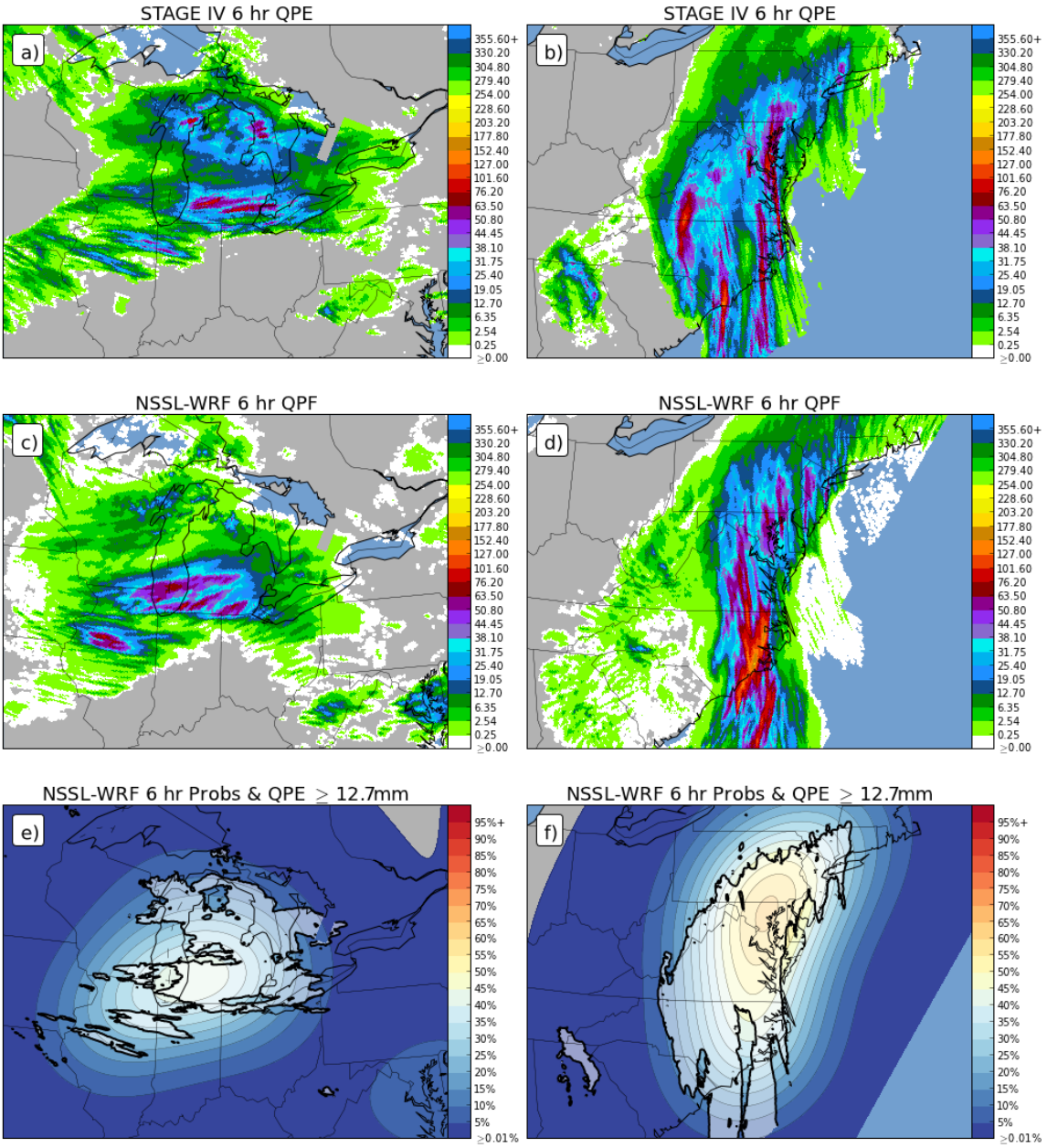

Figure 3.10: The same as in Figure 3.6 except using the 12.7 mm threshold.

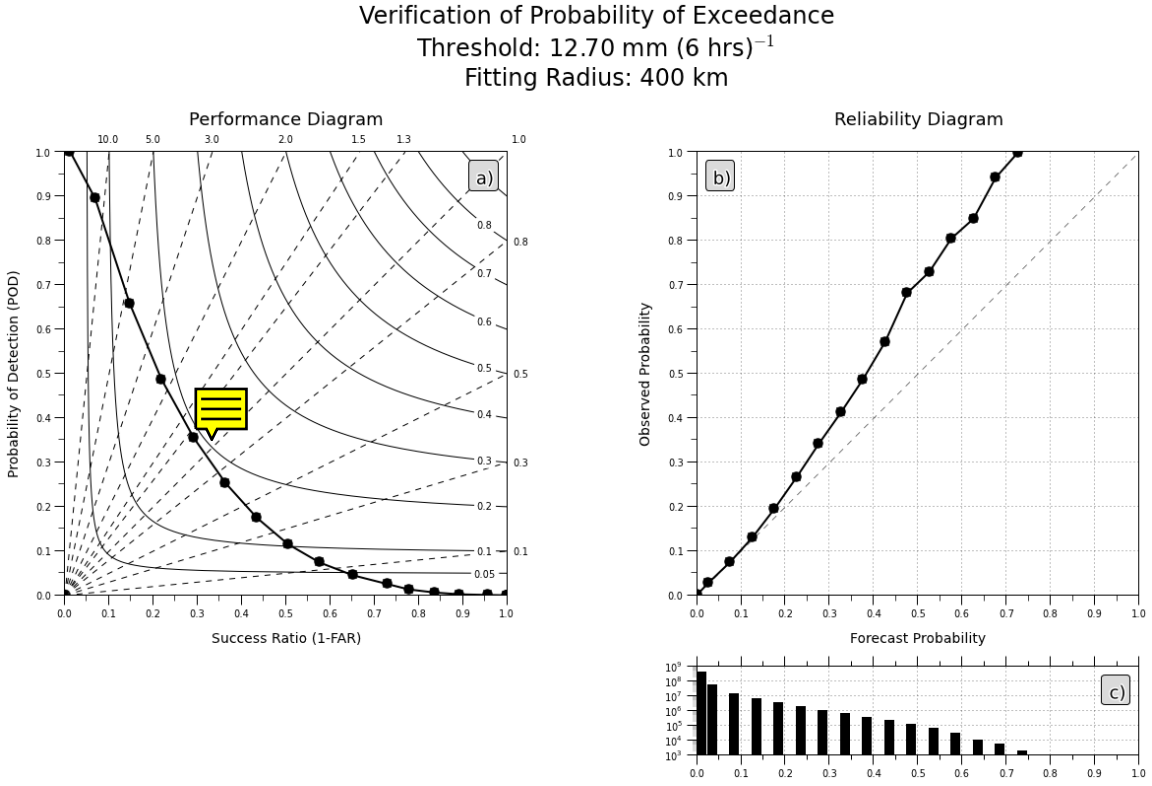

Figure 3.11: The same as in Figure 3.7 except using the 12.7 mm threshold.

Chapter 4

Ensemble

4.1 Ensemble Data

4.1.1 Ensemble Forecasts

As discussed at the beginning of chapter 3, in order to develop a statistical post-processing method for calibration of numerical weather prediction forecasts, both forecasts and observations must be readily available. Unfortunately, as mentioned in the Introduction, "...there is a limited database of forecasts for RCEs, making robust statistical techniques difficult..." As limited as databases of deterministic CAM forecasts are, the availability of *ensembles* of CAM forecasts is even worse.

Similar to CAM forecasts, most storm-scale ensemble forecast systems[1] (SSEF) have been erected on a temporary basis, produced in support of various field programs and experiments. The makeup of these SSEFs is often modified from one program to the next to adapt to the changing needs of the various experiments. Long running, consistent configured SSEFs were not available for this work. It is worth noting that as this work neared completion, the Air Force Weather Agency began running an operational SSEF.

One of the largest collections of SSEF forecasts has been produced by the University of Oklahoma's Center for the Analysis and Prediction of Storms (CAPS). Since 2007, CAPS has been producing CAM forecasts in support of the Hazardous Weather Testbed's

[1]The phrase storm-scale ensemble forecast is typically used in reference to ensembles composed of numerical forecasts produced without using cumulus parameterization.

(HWT) annual Spring Forecast Experiment (SFE). From 2007 through 2010, the ensemble configuration changed from year-to-year based on the results of previous years and updates to the WRF modeling system. In 2010, CAPS produced a 26 member SSEF. This SSEF was multi-model in nature, initialized at 00 UTC, used 4 km grid spacing, and integrated out to 30 hours. Of the 26 members, 19 were WRF-ARW[2] (Skamarock et al. 2008), 5 were WRF-NMM[3] (Rogers et al. 2009), and 2 were ARPS[4] (Xue et al. 2003). In 2011, CAPS expanded the SSEF from 26 members to 50, of which 41 were WRF-ARW, 5 were WRF-NMM, and 4 were ARPS. These forecasts were also initialized at 00 UTC, used a grid spacing of 4 km, but were integrated forward to 36 hours instead of 30.

For this thesis, fifteen members were chosen from the 2010 and 2011 ensembles because these members were configured almost identically between the two years. These 15 members are composed of 10 WRF-ARW forecasts, 4 WRF-NMM forecasts, and 1 ARPS forecast. The background initial conditions for these 15 members were downscaled from the 00 UTC 12 km NAM, with additional information coming from a three dimensional variational and cloud analysis from ARPS. Except for the control members (one each from the WRF-ARW, WRF-NMM, and ARPS), these initial conditions were then perturbed using mesoscale atmospheric perturbations from NOAA's Environmental Modeling Center's operational Short-Range Ensemble Forecast (SREF) system. Lateral boundary conditions for the three control members came from the 00 UTC NAM forecasts, whereas the remaining 12 members used the SREF forecast corresponding to the perturbations used in the initial conditions. A listing of the configurations for each member of the SSEF can be found in Table 4.1.

The only controlled change between 2010 and 2011 for the aforementioned set of 15 forecasts came from a change in the version of WRF. In 2010, WRF Version 3.1.1 was used whereas WRF Version 3.2.1 was used in 2011. Changes in the NAM and SREF,

[2]Weather Research and Forecasting – Advanced Research WRF
[3]Weather Research and Forecasting – nonhydrostatic Mesoscale Model
[4]Advanced Regional Prediction System

which were used for initial and lateral boundary conditions, were controlled by the NOAA National Weather Service and could not be avoided.

In 2010 the HWT SFE ran from 17 May through 18 June, and in 2011, the HWT SFE ran from 09 May through 10 June. CAPS provided SSEF forecasts each weekday during the experiment, with an additional two retrospective forecasts in 2011: one for the 27 April 2011 tornado outbreak in the southeast United States and another for the 22 May 2011 Joplin, Missouri EF-5 tornado. Table 4.2 lists the dates for which CAPS forecasts are available.

In an effort to be consistent with the evaluation carried out on the NSSL-WRF, the decision was made to use six-hour accumulation periods for the SSEF forecasts[5] and the Stage IV analyses. As was the case with the NSSL-WRF, the SSEF forecasts were drawn from subsets of the 12-36 hr forecast period ending at 18, 00, 06, and 12 UTC, but only for 6 hr time periods for which every member was available. Because the 2010 SSEF was only integrated out to 30 hr, the potential dataset was immediately reduced to 156 six-hour time periods (52 days times 3 periods a day).

4.1.2 Observations

As was the case for the evaluating the deterministic model calibration, NCEP's Stage IV national quantitative precipitation estimate analysis was chosen as the verification dataset. Please see section 3.1.2 for a description of the Stage IV dataset.

4.1.3 Processing

As stated in section 4.1.1, there were a maximum of 156 six-hour time periods for which SSEF data was expected. This potential dataset was further thinned by removing all time periods in which either one or more SSEF members were missing or the Stage IV analysis

[5]Unless otherwise noted, from this point forward the phrase "SSEF", "SSEF forecasts", or variants thereof refer to the 15 numerical forecasts that are presented in Table 4.1.

was unavailable. An additional 37 six-hour time periods were removed as a result, leaving 119 six-hour time periods, over 2010 and 2011, for which both the entire SSEF and the Stage IV analyses are available.

Next, as was necessary with the deterministic forecasts, the ensemble forecasts were interpolated to the Stage IV grid. All diagnostics and analyses were conducted on this grid. Furthermore, the same mask shown in Figure 3.3 was used to limit the analyses to areas east of the Rocky Mountains and near land.

Unfortunately, 119 time periods is a very small sample from which to evaluate a method for statical calibration of forecasts of rare events. This is especially true when the training and forecast data must both come from the 119 time periods. For comparison, the NSSL-WRF training dataset had 4120 training time periods and 1425 forecast time periods. In an attempt to maximize the utility of the 119 time periods, twenty "simulations" were created by resampling the 119 time periods. This was accomplished by randomly drawing, without replacement, 59 training periods from the valid 119 time periods with the remaining 60 periods used as forecast periods. From these 59 training periods, two-dimensional histograms were created for each of the 15 ensemble members. These histograms were then used to calculate the five fitting parameters — σ_x, σ_y, θ, h, and k — for each member's unique two-dimensional anisotropic Gaussian function. Forecasts for each member were then made for each time period of the 60 forecast time periods, with each member using its own fitted anisotropic Gaussian function. Thus, for each simulation, all 15 ensemble members used the same 59 time periods for training and the remaining 60 time periods for forecasting. This resulted in 18 000 forecasts (20 simulations x 60 forecast periods x 15 members).

4.2 Ensemble Results

4.2.1 25.4 mm Threshold

As previously mentioned, the first step of the proposed calibration process is to create two-dimensional composites of where observations occurred relative to forecast grid points. In this case that means creating two-dimensional composites for each member, for each of the twenty simulations. Visualizing, analyzing, and understanding all of these two-dimensional composites is challenging as there are 300 two-dimensional composites. To achieve this goal, a two-step approach was taken: 1) analyze each SSEF member's distribution for each the five Gaussian fitting parameters; and 2) examine aggregate measures of the spatial distributions.

A set of five figures (Figures 4.1–4.5) were created to examine the variability of each of the five Gaussian fitting parameters — one figure per fitting parameter. The figures contain box-and-whisker plots for each member's distribution of that figure's fitting parameter. The box-and-whisker plots offer insight into the variability in the various fitting functions used by each member. In each of the figures, the red horizontal line marks the median value of the distributions, and the blue box represents the 25th and 75th percentiles. The whiskers denote the range of the distribution, up to $\pm$ 1.5 times the interquartile range. Gold stars are used to denote any outliers, defined to be any value of the distribution that is outside the range of $\pm$ 1.5 times the interquartile range.

Figure 4.1 displays the box-and-whisker plots for the σ_x fitting parameter. This parameter is the length of the long axis of the fitted anisotropic Gaussian function. Nine out of the fifteen ensemble members do not have outliers. Of the remaining six ensemble members with outliers, two of them only have one outlier, two of them have two outliers, and one each has four and five outliers. Ensemble members that have outliers have a range of over 20 kilometers for σ_x, whereas those members without outliers generally have ranges of less

than 20 kilometers. This means that 40% of the ensemble members have variability in their σ_x parameter that was greater than 10% of the median length of the σ_x fitting parameter.

Figure 4.2 shows the box-and-whisker plots for the σ_y fitting parameter. This parameter is the length of the short axis of the fitted anisotropic Gaussian function. Unlike with the σ_x fitting parameter, the ensemble member distributions of the σ_y fitting parameter exhibit fewer outliers with only four members having them. This can be attributed to their being more variability in the 25th to 75th percentile, as noted by the increased size of the boxes for many members. Furthermore, whereas when the maximum range of the σ_x fitting parameter distribution greater than 20 kilometers indicated the presence of an outlier, several members have distributions of σ_y with a maximum range greater than 20 kilometers without having an outlier. In fact, six out of the ten WRF-ARW members have a range of σ_y greater than 20 kilometers and only two of them have outliers.

The distributions of the counter-clockwise rotation angle of the abscissa, fitting parameter θ, is shown in Figure 4.3. For this fitting parameter, eleven out of fifteen members exhibit outliers. This is not really surprising when one considers the overlap in the distributions of fitting parameters σ_x and σ_y. As alluded to in Section 2.2.2, when σ_x approaches σ_y, the Gaussian function approaches isotropy. As a Gaussian function approaches isotropy the variability of the rotation angle, θ increases as a result of the more circular nature to the function. It is much more difficult to accurately measure the orientation of the rotation angle of an ellipse that is nearly circular than that of one with a high level of eccentricity.

The large variability in the fitting parameters was not limited to just the shape of the anisotropic Gaussian. The location of the fitting Gaussian also ranged widely (fitting parameters h and k). Each member's distributions of the h fitting parameter (displacement in the east or west direction) is shown in Figure 4.4. Generally speaking, every SSEF member's distribution of h varies by over 20 kilometers (approximately 4-5 grid points), with many members demonstrating an even wider range, and that's just the variability within each member's distribution of h. Some of the values for h deviated from the forecast grid

point by over 40 kilometers, with the maximum difference between h values between any two members being over 80 kilometers!

Specifically, ARPS and WRF-NMM members consistently demonstrated a bias toward the observations centroid being displaced eastward, which corresponds to the forecasts, on average, being too far west. Most of the WRF-ARW members exhibited the opposite behavior, namely having the observations too far westward (indicating a eastward bias with the forecast). Not every WRF-ARW member exhibited this bias, however. Two of the WRF-ARW members have a majority of the distribution consist of positive values for h indicating a westward forecast bias. However, unlike with the ARPS and WRF-NMM members where the entire distribution consisted of positive values for h, every member of the WRF-ARW had at least a portion of the distribution with negative values, indicating an eastward forecast bias compared to observations.

Similar variability in the distributions of fitting parameter k, displacement in the north-south direction, are also observed (Figure 4.5). One difference, however, is that there appears to be a forecast displacement preference for most members. Most members' forecasts appear to be farther north than the centroid of observations.

In addition to examining the two-dimensional composites in the context of their one-dimensional distributions of the five anisotropic Gaussian fitting parameters, a cursory examination of their spatial characteristics was also conducted. The standard deviation of each member's two-dimensional composites is shown in Figure 4.6. A quick examination of the standard deviation of the two-dimensional composites does not yield any immediate insights. Some members exhibit a maximum in variability along a southwest to northeast orientation, whereas other members exhibit a more uniform decrease in variability as one moves radially outward from the forecast grid point. In both of these orientations, the maximum variability tended to be located near the forecast grid point, with generally decreasing variability as distance from the forecast grid point increases.

Unfortunately, Figure 4.6 doesn't yield much insight into what the specific two-

dimensional composites used in the various simulations might look like. To examine that aspect of the distributions, a single simulation was chosen at random from the twenty simulations to examine further. Figure 4.7 shows the two-dimensional composites for each member at the 25.4 mm in 6 hr threshold. In this simulation, it is easily suggested that the WRF-NMM ensemble members are generally the wettest, as indicated by the substantially more observations, with the WRF-ARW members generally drier. (The ARPS member is in between.) The overall axis of observations relative to forecasts indicates a general southeast-to-northeast orientation in most members, although this is more readily apparent in some members than others.

The lower right panel of Figure 4.6 depicts the standard deviation between the two-dimensional composites of each member for all simulations. In this panel, the darker colors indicate a greater standard deviation at that particular grid point than grid points with a lighter color. There is considerable variability over the southern and eastern portion of the composite radius, with the maximum in variability at relatively far distances to the south, and slightly east. The least variable portion region over the compositing radius is found to the north and northwest.

Although the orientation of the axis of maximum observations tends to generally be similar between members, the centroid of the distribution exhibits more variability. About half of the members have the centroid of observations too far east and half being too far west. (This corresponds to the h parameter of the Gaussian fitting parameters.) The members having the observations centroid too far east tended to be farther away from the forecast than those members having the observations centroid too far west. In this simulation, the WRF-NMM has a westward bias with its forecasts, as the centroid of observations is east of the forecast grid point in every WRF-NMM member. A similar observation cannot be made for the WRF-ARW members. It is unclear from this single simulation if the westward forecast bias in the WRF-NMM members is systematic of the WRF-NMM core, or merely a function of only having 4 WRF-NMM members.

When examining the north-south variations of the observations centroid relative to the model forecast, similar variations are observed (not shown). (The north-south displacement of the observations centroid corresponds to the k parameter of the Gaussian fitting parameters.) Slightly more than half the members have the centroid of observations too far north, indicating a southward bias of the forecast, with slightly less than half having the centroid of observations too far south. Three of the four WRF-NMM members had the forecast too far north, with the WRF-NMM control member having the greatest displacement. No obvious preference in displacement direction is readily apparent from the WRF-ARW members. However, it does appear that generally speaking, the magnitude of the displacement of the WRF-ARW members appears to be less than that of the WRF-NMM members.

Similar to the lower right panel of Figure 4.6, the lower right panel of Figure 4.7 depicts the standard deviation between the two-dimensional composites of each member for the given simulation. In this panel, the darker colors indicate a greater standard deviation at that particular grid point than grid points with a lighter color. It is readily apparent that the greatest variability between the members exists to the east of the forecast grid point, as was the case with the overall variability denoted in the lower right panel of Figure 4.6. This is due to the wetter WRF-NMM members having a westward forecast bias, resulting in a wider range of observation counts to the east of the forecast.

Although there appears to be variability in the various anisotropic Gaussian functions used to model the displacement characteristics of the SSEF members, how does the modeled Gaussian work in practice? Figure 4.8 shows the probabilistic forecast from each member of the SSEF for the six hours ending 06 UTC 20 May 2010. Meteorologically speaking, this time period comes at end of a severe weather day; the SPC issued a High Risk for severe weather across portions of central Oklahoma for much of the day 19 May 2010. Thunderstorms developed across northwestern Oklahoma and southern Kansas early in the afternoon and developed eastward from there. Several clusters of thunderstorms emerged through the course of the afternoon, culminating in a heavy rain event across por-

tions of eastern Oklahoma, eastern Kansas, western Missouri, and western Arkansas. A large area of observed precipitation amounts exceeding the 25.4 mm in 6 hr threshold occurred across eastern Oklahoma into western Arkansas and southwestern Missouri, with a smaller area observed across northeast Kansas into northwest Missouri. Additionally, a small area exceeding the threshold was observed in southwest Kansas to the north of Dodge City, KS.

In terms of the overall appearance of their probabilistic forecasts, each member appears to produce a probabilistic forecast that is maximized in the vicinity of the largest area of observed precipitation accumulation, with decreasing probabilities to the north and south. Furthermore, every member of the SSEF seems to have captured the general area where 25.4 mm in 6 hr occurred. In fact, within the limited domain shown in Figure 4.8, all observed areas exceeding the given threshold were captured by non-zero probabilities, with most areas, for most SSEF members, captured within at least a 5% contour. However, no members' probabilistic forecast encapsulated all observed areas with at least a 5% contour. Although each probabilistic forecast appears to highlight the same area for a heavy rain potential, the character of each member's probabilistic forecast is different. All members appear to convey the threat of an elongated north-south area with the potential to see heavy rain, but a handful of members also suggest the possibility of the heavy rain extending northwestward into central and western Kansas. The ARPS member and the WRF-ARW control member both appear to have their highest probabilities farther north than the other members, but still south of the secondary area of observed 25.4 mm in 6 hr.

The lower right panel of Figure 4.8 is the "ensemble" probability forecast. It was computed by taking the average of all fifteen SSEF probabilities at each grid point. Because each of the SSEF members highlight the same general location with their individual forecasts, the ensemble average appears to be a good forecast as well. However, one drawback of an ensemble average forecast is that the potential exists for the resulting forecast to be overly smooth. This is because averaging tends to dampen the higher probabilities from

each of the SSEF members due to the low likelihood of the maximum probabilities from each member actually occurring on the same grid point in all of the members. Furthermore, if the individual probability forecasts were spatially very different, the ensemble average field would essentially average out all of the signal offered by the individual members.

As previously mentioned, it is incorrect to assess the validity of a probabilistic forecast on the basis of a single probabilistic forecast. Instead, probabilistic forecasts must be evaluated over a sufficiently large sample of forecasts, such that each probability forecast is used a sufficient number of times to be accurately evaluated. In an idealized world, probabilistic forecasts would be evaluated over an infinite number of forecasts. Since an infinite number of forecasts is impossible, it is generally understood that the larger the sample of probabilistic forecasts to be evaluated, the more robust the verification of the probabilities will be.

Unfortunately, with only 60 forecast time periods for each of the SSEF members, and dealing with a rare event, drawing meaningful verification statistics is a challenge. The approach used here was to create performance diagrams and reliability diagrams for each member (Figure 4.9). Rather than plot the curves of all twenty simulations for all fifteen members to visualize the uncertainty in the verification, only the mean curve for each member is plotted. In the case of a reliability diagram, visualizing the uncertainty of the mean reliability is easily achieved by plotting error bars of observed probability around the mean reliability at each members' unique forecast probability. Here, the error bars, or uncertainty in the mean, were calculated to be $\pm$ one standard deviation of the observed probabilities of that member's forecast probability at that given probability threshold. In an effort to prevent too many lines from cluttering the reliability diagram, instead of using error bars the area bounded by $\pm$ one standard deviation of each member's mean reliability is color filled with a semi-transparent color.)

Visualizing uncertainty on a performance diagram is more difficult. The plotted point is based upon that forecast probability's Success Ratio (abscissa) and Probability of De-

tection (ordinate). Since both of these are derived verification measures each can exhibit variability. In other words, the location of each evaluated forecast probability is not fixed along either the abscissa or the ordinate. Furthermore, instead of being able to add a single set of error bars, error bars must be added in both coordinate directions, making for a cluttered figure. Once again, instead of adding error bars, which would make the figure even more difficult to read, an ellipse is drawn around each data point, where the ellipse's x-axis is given by the standard deviation in the abscissa and the ellipse's y-axis is given by the standard deviation in the ordinate.

Figure 4.9 depicts both the performance diagram (panel a) and the reliability diagram (panel b) for the SSEF ensemble at the 25.4 mm threshold. Looking at the performance diagram it is readily apparent that the overall values appear to be worse than those from the NSSL-WRF in Figure 3.7, with CSI scores being less than 0.1 for all probability thresholds from all ensemble members. Visually, it does not appear there is much difference between the individual SSEF member forecasts (black curves) and the ensemble forecast generated by a modified Hamill and Colucci method (red curve), as the latter forecasts appear to fall within the spread of SSEF member forecasts. It does appear as if there is a slight improvement by using the ensemble mean forecast (blue curve), but it does not offer much improvement. Looking at the "error clouds" associated with each point, it does not look like there is much variability in the performance diagram plots between each simulation for each SSEF member.

Examining the reliability diagrams found in Figure 4.9b yields a slightly different story. While the performance diagram suggests the forecast "goodness" was generally poor, the reliabilities appear to be fairly good for such a rare event and such a small sample size. By no means is the reliability near perfect for the SSEF member forecasts, but except for the rarer probability forecasts of some SSEF members (i.e., higher probability forecasts), every member of the SSEF falls between perfect reliability and the line of no skill (in a Brier Score sense; not shown). The variability of the reliability diagram to changes in

simulation appears to be greater at higher forecast probabilities, of which their are fewer forecast grid points to evaluate. This is easily seen in the general increase in width of the error plumes surrounding the mean reliability as the forecast probabilities increase. It should also be noted that, except for the rare probability threshold forecasts, it appears that every member of the SSEF is closer to perfect reliability than the forecasts generated by the modified Hamill and Colucci method. However, one benefit of the modified Hamill and Colucci method forecasts is that probability forecasts occur over a wider range of forecast thresholds. In other words, it is possible to have a near 100% probability forecast, that is currently not possible from the individual SSEF member forecasts.

One thing of note in the reliability diagrams is that every probabilistic forecast, for every member, including the ensemble mean (blue curve) and modified Hamill and Colucci generated forecasts, is an over forecast. In other words, the number of grid point observations at a given forecast probability is fewer than the number of observations needed to achieve perfect reliability for that particular member. One interesting artifact of the over forecasts by every member of the SSEF is that the ensemble average probability forecasts appear to be closer to perfect reliability than any of the individual SSEF members. This is because the averaging process acts as a spatial smoother, damping the peak probabilities, and spreading them out over a slightly larger area. Averaging a set of probability forecasts has the affect of shifting the resulting reliability curve to the left, as compared to the collection of reliability curves for each individual member. Thus, because every SSEF member over forecasts, the resulting damped, and spatially smoothed forecasts caused by averaging will be closer to perfect reliability than the individual forecasts[6].

[6]This is not always guaranteed. In the limiting case of every ensemble member producing the same exact forecast, the ensemble average will be identical to the collection of ensemble forecasts.

4.2.2 12.7 mm Threshold

As was the case with the NSSL-WRF, the analysis of the proposed calibration method was also carried out at the 12.7 mm in 6 hr threshold using the SSEF data. This was partly motivated by the high variability in both the two-dimensional composites and the resulting reliability curves. By examining the proposed technique at a lower precipitation accumulation threshold, it is hoped that more forecast events (grid points) will occur. The increase in precipitation events would allow for more filled out two-dimensional composites, and hopefully decrease the intra member variability in the two-dimensional composites between simulations. This should, in turn, provide for a larger sample from which to conduct the verification and evaluate the probabilistic forecasts. The format here follows the same as with the 25.4 mm in 6 hr threshold discussed in Section 4.2.1.

As was the case before, a set of five figures (Figures 4.10–4.14) were created to examine the variability of each of the five Gaussian fitting parameters — one figure per fitting parameter. Figure 4.1 displays the box-and-whisker plots for the σ_x fitting parameter. This parameter is the length of the long axis of the fitted anisotropic Gaussian function. Fairly important differences arise when comparing the distributions of the σ_x fitting parameter between the 25.4 mm threshold and the 12.7 mm threshold. In the case of the 25.4 mm in 6 hr threshold, six members had distributions in which the range of σ_x values exceeded 20 kilometers. With the lower threshold of 12.7 mm in 6 hr the range between the largest and smallest values of σ_x is less than 18 kilometers. In fact, most members have a range of σ_x values that is less than 10 kilometers.

Similar results are found when examining the box-and-whisker plots for the σ_y fitting parameter (Figure 4.2) as were found with the σ_x distributions. Once again the maximum range between any two σ_y values from any member is approximately 15 kilometers, which is about half the maximum difference between any two σ_y values at the 25.4 mm threshold; most members exhibit a range of less than 10 kilometers. Furthermore, whereas the distri-

butions of σ_x and σ_y had quite a bit of overlap at the 25.4 mm threshold, there is no overlap between the two fitting parameters at the 12.7 mm threshold. It is worth noting that six of the fifteen members exhibit an outlier value for σ_y at the 12.7 mm threshold, but even those members exhibit a range of less than 15 kilometers.

The distributions of the counter-clockwise rotation angle of the abscissa, fitting parameter θ, is shown in Figure 4.12. For this fitting parameter, six out of fifteen members exhibit outliers, which is nearly half the number of members with outliers at the 25.4 mm threshold. This is not really surprising when one considers the lack of any overlap in the distributions of fitting parameters σ_x and σ_y at the 12.7 mm threshold. Recall when σ_x approaches σ_y, the Gaussian function approaches isotropy, resulting in nearly circular distributions that make it difficult to assess the angle of rotation. With no overlap in the distributions of σ_x and σ_y, this ensures that the resulting two-dimensional composites of some degree of eccentricity and a somewhat better defined abscissa rotation angle. Even the SSEF members that do have outliers, the range of θ values is still significantly less than the variability seen in θ at the 25.4 mm threshold. In fact, the median rotation angle (θ) from all of the members appears to be within 5-7 degrees of 45°, with the maximum orientation being slightly more than 70°and the minimum orientation being slightly less than 30°. This means that the fitted anisotropic Gaussian for all two-dimensional composites at the 12.7 mm threshold exhibits some variation on a southwest-to-northeast orientation.

The improvement in the variability in the fitting parameters was not limited to just the shape of the fitted anisotropic Gaussian. The variability in the location of the fitting Gaussian also decreased (fitting parameters h and k), although not by as much as with σ_x, σ_y, and θ. Each member's distributions of the h fitting parameter (displacement in the east or west direction) is shown in Figure 4.13. Generally speaking, every SSEF member's distribution of h varies by near, or slightly less than, 20 kilometers (approximately 4 grid points).

Specifically, ARPS and WRF-NMM members consistently demonstrated a bias toward

the observations centroid being east, which corresponds to the forecasts, on average, being too far west. Most of the WRF-ARW members exhibited the opposite behavior, namely having the observations too far west (indicating an eastward bias with the forecast). Not every WRF-ARW member exhibited this bias, however. Two of the WRF-ARW members have a majority of the distribution consist of positive values for h indicating a westward forecast bias. However, unlike with the ARPS and WRF-NMM members where the entire distribution consisted of positive values for h, every member of the WRF-ARW had at least a portion of the distribution with negative values, indicating an eastward forecast bias compared to observations. Similar variability in the distributions of fitting parameter k, displacement in the north-south direction, are also observed (Figure 4.14). One difference, however, is that there appears to be a forecast displacement preference for most members. Most members' forecasts appear to be farther north than the centroid of observations. These results are generally consistent with the distributions of h and k at the 25.4 mm threshold, albeit with slightly more restricted distributions.

Next, the two-dimensional composites spatial characteristics are examined. The standard deviation of each member's two-dimensional composites is shown in Figure 4.15. A quick examination of the standard deviation of the two-dimensional composites shows that the two-dimensional composites appear to be more uniform than those for the 25.4 mm threshold. In other words, the general pattern to the standard deviations suggests that there is not any obvious spatial biases. It does appear that the WRF-NMM members have more variability to the east of the representative forecast grid point, and that there may be more of a maximum in variability from the southwest-to-northeast for these members, but is not as obvious as shown in Figure 4.6. In all standard deviation plots the maximum variability tends to be located near the forecast grid point, with generally decreasing variability as distance from the forecast grid point increases.

The lower right panel of Figure 4.15 depicts the standard deviation between the two-dimensional composites of each member for all simulations. In this panel, the darker colors

indicate a greater standard deviation at that particular grid point than grid points with a lighter color. There is considerable variability over the southeastern and eastern portion of the composite radius, with the maximum in variability at relatively far distances to the south, and slightly east. The least variable portion region over the compositing radius is found to the north, south, and west, with decreasing variability the farther away from the forecast grid point.

Unfortunately, Figure 4.15 doesn't yield much insight into what the specific two-dimensional composites used in the various simulations might look like. To examine that aspect of the distributions, the same simulation used in Figure 4.7 was used to examine further. Figure 4.16 shows the two-dimensional composites for each member at the 12.7 mm in 6 hr threshold. In this simulation, it is easily suggested that the WRF-NMM ensemble members are generally the wettest, as indicated by the increased number of observations, with the WRF-ARW members generally drier. (The ARPS member is still in between.) The overall axis of observations relative to forecasts indicates a general southeast-to-northeast orientation in most members, although this is more readily apparent in some members than others.

The lower right panel of Figure 4.15 depicts the standard deviation between the two-dimensional composites of each member for all simulations. In this panel, the darker colors indicate a greater standard deviation at that particular grid point than grid points with a lighter color. There is considerable variability over the southern and eastern portion of the composite radius, with the maximum in variability at relatively far distances to the south, and slightly east. The least variable portion region over the compositing radius is found to the north, west, and far south.

Although the orientation of the axis of maximum observations tends to generally be similar between members, the centroid of the distribution exhibits more variability. About two-thirds of the members have the centroid of observations too far east and one-third have the centroid be too far west. (This corresponds to the h parameter of the Gaussian fitting

parameters.) This represents an increase in the number of members where the centroid of observations is found to be to the east. The members having the observations centroid too far east tended to be farther away from the forecast than those members having the observations centroid too far west. This is consistent with the findings at the 25.4 mm threshold, although the gap between the two groups of members has decreased. In this simulation, the WRF-NMM has a westward bias with its forecasts, as the centroid of observations is east of the forecast grid point in every WRF-NMM member. A similar observation cannot be made for the WRF-ARW members. As was also the case at the 25.4 mm threshold, it is unclear from this single simulation if the westward forecast bias in the WRF-NMM members is systematic of the WRF-NMM core, or merely a function of only having 4 WRF-NMM members.

When examining the north-south variations of the observations centroid relative to the model forecast, similar variations are observed. (The north-south displacement of the observations centroid corresponds to the k parameter of the Gaussian fitting parameters.) Slightly more than half the members have the centroid of observations too far south, indicating a northward bias of the forecast, with slightly less than half having the centroid of observations too far north. This is the opposite that at the 25.4 mm threshold. Three of the four WRF-NMM members had the forecast too far north, with the WRF-NMM S4M4 member having the least displacement. No obvious preference in displacement direction is readily apparent from the WRF-ARW members. However, it does appear that generally speaking, the magnitude of the displacement of the WRF-ARW members appears to be less than that of the WRF-NMM members.

Similar to the lower right panel of Figure 4.15, the lower right panel of Figure 4.16 depicts the standard deviation between the two-dimensional composites of each member for the given simulation. In this panel, the darker colors indicate a greater standard deviation at that particular grid point than grid points with a lighter color. It is readily apparent that the greatest variability between the members exists to the east of the forecast grid point,

as was the case with the overall variability denoted in the lower right panel of Figure 4.15. This is due to the wetter WRF-NMM members having a westward forecast bias, resulting in a wider range of observation counts to the east of the forecast.

How does the modeled Gaussian work in practice? Figure 4.17 shows the probabilistic forecast from each member of the SSEF for the six hours ending 06 UTC 20 May 2010, the same as in Figure 4.8. At the lower precipitation accumulation threshold of 12.7 mm in 6 hr, a large area of observed precipitation amounts exceeding the threshold occurred across eastern Oklahoma and western Arkansas and extended north and northeastward into western and central Missouri as well as eastern and northeastern Kansas. Additionally, small areas exceeding the threshold were observed in northeast Louisiana and extreme southeastern Arkansas, south-central into southwestern Kansas, northern Texas Panhandle, eastern Colorado, and southeastern Wyoming.

In terms of the overall appearance, each member appears to produce a probabilistic forecast that is maximized in the vicinity of the largest area observed precipitation accumulation exceeding the specified threshold, with decreasing probabilities to the north and south. Furthermore, every member of the SSEF seems to have captured the general area where 12.7 mm in 6 hr occurred. In fact, as was the case at the higher precipitation threshold, within the limited domain shown in Figure 4.17, all observed areas exceeding the given threshold were captured by non-zero probabilities, with most areas, for most SSEF members, captured within at least a 5% contour. However, no members' probabilistic forecast encapsulated all observed areas with at least a 5% contour. Although each probabilistic forecast appears to highlight the same area for a heavy rain potential, the character of each member's probabilistic forecast is different. All members appear to convey the threat of an elongated north-south area with the potential to see heavy rain, with most members also suggesting the possibility of the heavy rain to extend northwestward into central and western Kansas. None of the members captured the observed area in northeast Louisiana, nor the areas in the northern Texas Panhandle and southeastern Wyoming. It is tempting to say

the ensemble forecasts are poor in these areas, however, that cannot be determined based on the figures shown. This is because the calibration approach presented here requires a large number of nearby grid points forecast to exceed the specified threshold in order produce a probabilistic forecast that exceeds the 5% contour. It is quite possible that the individual SSEF members did accurately predict accumulations in these "missed" areas that exceeded the 12.7 mm in 6 hr threshold, but did so over a small area which yielded small forecast probabilities.

Lastly, the lower right panel of Figure 4.17 is the "ensemble" probability forecast. Because each of the SSEF members highlight the same general location with their individual forecasts, the ensemble average appears to be a good forecast also.

Figure 4.18 displays the performance and reliability diagram for the 12.7 mm in 6 hr threshold. Looking at the performance diagram (Figure 4.18a) it can be seen that most curves have shifted to the upper-right, indicating improved forecasts as compared to the 25.4 mm forecasts. As was the case at the great accumulation threshold, there is little variability in the curves for an individual member, as indicated by the small "error cloud" surrounding each plotted point. The error clouds do increase as one moves left-to-right along the curve (i.e., as one moves from lower forecast probability to higher forecast probability), which is expected as the higher forecast probabilities are forecasted fewer times. Once again the forecast probabilities generated by the modified Hamill and Colucci method (red curve) are shown to perform slightly better over a larger range of forecast probabilities as compared to the individual SSEF member forecasts. The SSEF ensemble generated probabilities (blue curve) are similar to those from the modified Hamill and Colucci method over most of the diagram, but perform slightly better at some of the higher forecast probabilities.

Examining the reliability diagram (Figure 4.18b) much better results are shown as compared to the reliability diagrams constructed for the 25.4 mm forecasts. For all SSEF forecasts, the range over which probability forecasts are generated has increased from a maximum of just under 50% for the higher members at the 25.4 mm threshold to just under

the 70% at the 12.7 mm threshold. It's also readily seen that the SSEF member reliability curves are much closer to perfect reliability than before. In fact, with the exception of each member's highest value probability forecasts (right most portions of the black curves), the SSEF member's reliability is generally within 10% of the observed probabilities.

As was the case at the 25.4 mm threshold, the forecast probabilities generated by the modified Hamill and Colucci method are the worst performing forecasts, strictly in terms of reliability. Over most probability forecasts, the reliability curve for the modified Hamill and Colucci forecasts is well separated from the reliability curves constructed from the probabilistic forecasts generated directly from each ensemble member. One thing that is import to note, is that unlike what occurred with the forecasts at the 25.4 mm threshold, the ensemble averaged probabilities are *less* reliable than the individual SSEF member forecasts. Upon reflection, this is not surprising. As explained toward the end of Section 3.2.1, the act of averaging the ensemble forecasts results in a reliability curve being shifted to the left as compared to the reliability curves constructed from the individual members that make up the ensemble. Since the reliability curves for each SSEF member are near perfect reliability to begin with, constructing a forecast from the average forecast probabilities will result in that forecast's reliability curve to shift to the left and under forecast the observed probability at each forecast probability.

4.2.3 Discussion

It is readily apparent from examining Figures 4.1–4.7 and 4.10–4.16 that there is a wide range of variability amongst the fitting parameters at all precipitation accumulation thresholds. Furthermore, examining the verification plots (Figures 4.9a,b and 4.18a,b) could lead one to believe that the probabilistic forecasts are not very good. However, it is important to take these results in proper context.

Warm Season verus Cool Season

The period from late Spring through Summer is traditionally the hardest time for numerical weather prediction with respect to quantitative precipitation forecasts! To illustrate this point, the 12.7 mm (0.5 in) in 6 hr quantitative precipitation forecast verification scores of the NAM and Global Forecast System (GFS) numerical models, as well as the human forecasts produced by the NOAA Hydrometeorological Prediction Center (HPC; Available online at `http://www.hpc.ncep.noaa.gov/html/hpcverif.shtml#6hour`) are shown in Figures 4.19–4.22 . The figures show the threat score (equivalent to CSI) of each forecast system plotted as a function of month for the same forecast time periods as used in both the NSSL-WRF and the SSEF. The general trend is for a decline in forecast verification scores during the warm season (late Spring through Summer) and an increase in the cool season (late Autumn through Winter). Although these verification scores are for the year 2012, a different time period than used in the analyses conducted in this paper and are only valid at a single precipitation accumulation threshold, the general trend is similar to previous years and higher accumulation thresholds.

The evaluation of the calibration method, as applied to the SSEF, has been conducted at what is traditionally one of the most challenging times for quantitative precipitation forecasting, both for numerical weather prediction and for humans! Bearing this in mind, a closer look at the performance diagram of the 12.7 mm in 6 hr threshold indicates that the CSI scores at the various probabilistic forecast values for the SSEF members are slightly better than the CSI values from the larger-scale numerical models (GFS and NAM) for the corresponding time of year. This means that it is possible to construct a deterministic forecast from the probabilistic forecast of any SSEF member that is more skillful than the GFS or NAM. This is achieved for any SSEF member by choosing a forecast probability threshold for that member that exhibits a higher CSI than either the NAM or GFS and converting all forecast probabilities less than that threshold into a "NO" forecast and forecast prob-

abilities exceeding that threshold into a "YES" forecast. Here a "NO" forecast indicates that the grid point is not expected to exceed the specified threshold, and a "YES" forecast indicates that the grid point is expected to exceed the specified threshold. In fact, for some members of the SSEF, it would be possible to generate a deterministic forecast from the probabilistic forecast that would be comparable in skill to the human forecasts produced by HPC[7].

It is important to note that this more skillful deterministic forecast can be created from any of the SSEF members. In other words, it is possible to produce a collection of deterministic forecasts, derived from probabilistic forecasts, that are more skillful than the larger-scale numerical models. The fact the more skillful deterministic forecasts can be derived from the probabilistic forecasts is an important one. This allows for improvements in the current deterministic paradigm in which a majority of end users operate, but it also provides additional useful information for higher-end users. End users who have very specific cost-loss thresholds can utilize the probabilistic forecast to gain additional, user-specific information from the probabilistic forecasts, without negatively impacting users who depend on the legacy deterministic forecasts.

Limited Training Period

With only 59 time periods from which to try and capture each SSEF member's displacement characteristics, achieving the threshold of 25.4 mm in 6 hr occurs too infrequently to create a two-dimensional composite from which to calculate robust fitting parameters. The 59 training time periods is simply not long enough to adequately capture the variability in each member's spatial displacement for such a rare event. For instance, the maximum number of observations at any of the two-dimensional composite grid points, from any SSEF member, is less than 25 000. Compare that with the hundreds of thousands found

[7]HPC does not verify their quantitative precipitation forecasts on a grid point by grid point basis. Instead they verify at a number of predefined locations. Please see Olsen et al. (1995) for additional information on how HPC conducts its verification.

in the NSSL-WRF composite at the same threshold. This inability to capture all of the displacement variability for all SSEF members significantly undercuts the justification of the proposed calibration method. Namely, that by modeling the displacement errors, and using the modeled displacement errors as the basis of KDE, one can create calibrated probabilistic forecasts from deterministic models. Unfortunately, the method is only as good as the ability to correctly model the underlying displacement errors. If the resulting two-dimensional composite is unable to adequately represent the displacement error, the method will not produce as well of a forecast as it could.

In the case of the SSEF, simply moving to a lower threshold, such as the 12.7 mm in 6 hr, resulted in the maximum number of observations at any two-dimensional composite grid point being greater than one hundred thousand – even for such a limited time period! This allowed for an a better depiction of the underlying displacement error, and a much more reliable probabilistic forecast was achieved.

Table 4.1: Configurations for the 2010 and 2011 CAPS Ensemble Members.

Member	I. C.	B. C.	Radar	Micro-physics	LSM	PBL
arw_cn	00Z ARPS 3DVAR & Cloud Analysis	00Z NAM Forecast	Yes	Thompson	Noah	MYJ
arw_m4	arw_cn + em_p1_pert	21Z SREF em_p1	Yes	Morrison	RUC	YSU
arw_m5	arw_cn + em_p2_pert	21Z SREF em_p2	Yes	Thompson	Noah	QNSE
arw_m6	arw_cn - nmm_p1_pert	21 SREF nmm_p1	Yes	WSM6	RUC	QNSE
arw_m7	arw_cn + nmm_p2_pert	21Z SREF nm_p2	Yes	WDM6	Noah	MYNN
arw_m8	arw_cn + rsm_n1_pert	21Z SREF rsm_n1	Yes	Ferrier	RUC	YSU
arw_m9	arw_cn - etaKF_n1_pert	21Z SREF etaKF_n1	Yes	Ferrier	Noah	YSU
arw_m10	arw_cn + etaKF_p1_pert	21Z SREF etaKF_p1	Yes	WDM6	Noah	QNSE
arw_m11	arw_cn - etaBMJ_p1_pert	21Z SREF etaBMJ_p1	Yes	WSM6	RUC	MYNN
arw_m12	arw_cn + etaBMJ_p1_pert	21Z SREF etaBMJ_p1	Yes	Thompson	RUC	MYNN
nmm_cn	00Z ARPS 3DVAR & Cloud Analysis	00Z NAM Forecast	Yes	Ferrier	Noah	MYJ
nmm_m3	nmm_cn + nmm_n1_pert	21Z SREF nmm_n1	Yes	Thompson	Noah	MYJ

Continued on Next Page . . .

Table 4.1 – Continued

Member	I. C.	B. C.	Radar	Micro-physics	LSM	PBL
nmm_m4	arw_cn + nmm_n2_pert	21Z SREF nmm_n2	Yes	WSM6	RUC	MYJ
nmm_m5	arw_cn + em_n1_pert	21Z SREF em_n1	Yes	Ferrier	RUC	MYJ
arps_cn	00Z ARPS 3DVAR & Cloud Analysis	00Z NAM Forecast	Yes	Lin	Force Restore	1.5-order TKE-based

Note 1: For all members, cumulus parameterization is turned off

Note 2: For all ARW members, the long-wave radiation parameterization is RRTM and the short-wave radiation parameterization is Goddard

Note 3: For nmm_cn, nmm_m2, & nmm_m3 the long-wave radiation parameterization is GFDL and the short-wave radiation parameterization is GFDL

Note 4: For nmm_m4 & nmm_m5 the long-wave radiation parameterization is RRTM and the short-wave radiation parameterization is Dudhia

Note 5: The arps member uses Chou/Suarex for radiation

Note 6: Ferrier+ refers to a subset of changes in the updated version now in NEMS/NMMB

Note 7: The ARPS PBL scheme (Xue et al. 1996; Sun and Chang 1986) uses a non-local vertical mixing length within the convective boundary layer

Table 4.2: 2010 and 2011 dates where CAPS forecasts are available.

2010 Dates	
Week #1:	17–21 May 2010
Week #2:	24–28 May 2010
Week #3:	31 May – 04 June 2010
Week #4:	07–11 June 2010
Week #5:	14–18 June 2010
2011 Dates	
Week #1:	09–13 May 2011
Week #2:	16–20 May 2011
Week #3:	23–27 May 2011
Week #4:	30 May – 03 June 2011
Week #5:	06–10 June 2011
Special Run Dates	
Special #1:	27 April 2011
Special #2:	22 May 2011

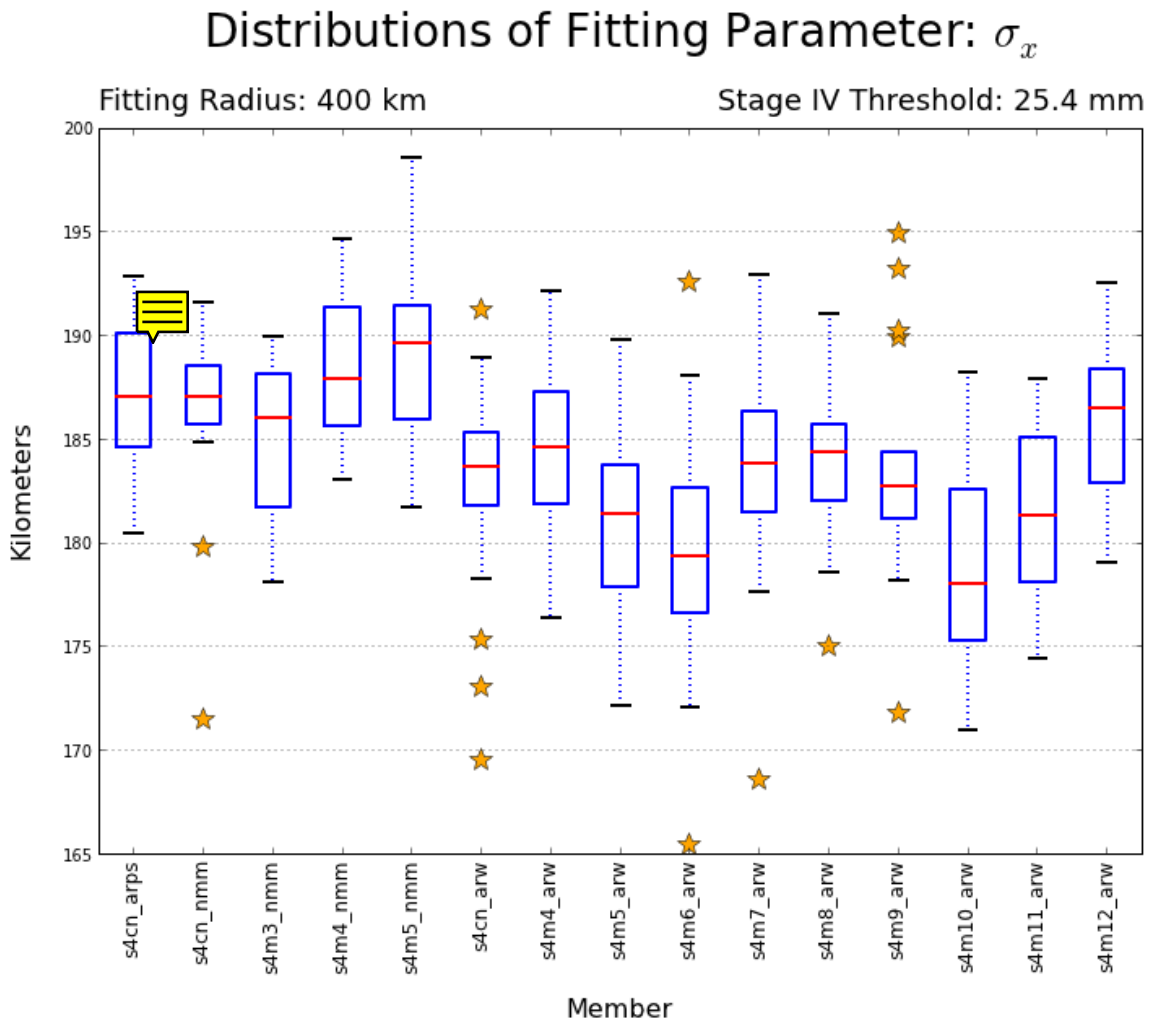

Figure 4.1: Box-and-Whisker plots of the σ_x anisotropic Gaussian fitting parameter, for each member of the SSEF at the 25.4 mm in 6 hr threshold. Each SSEF member's distribution is derived from the twenty re-sampled simulations.

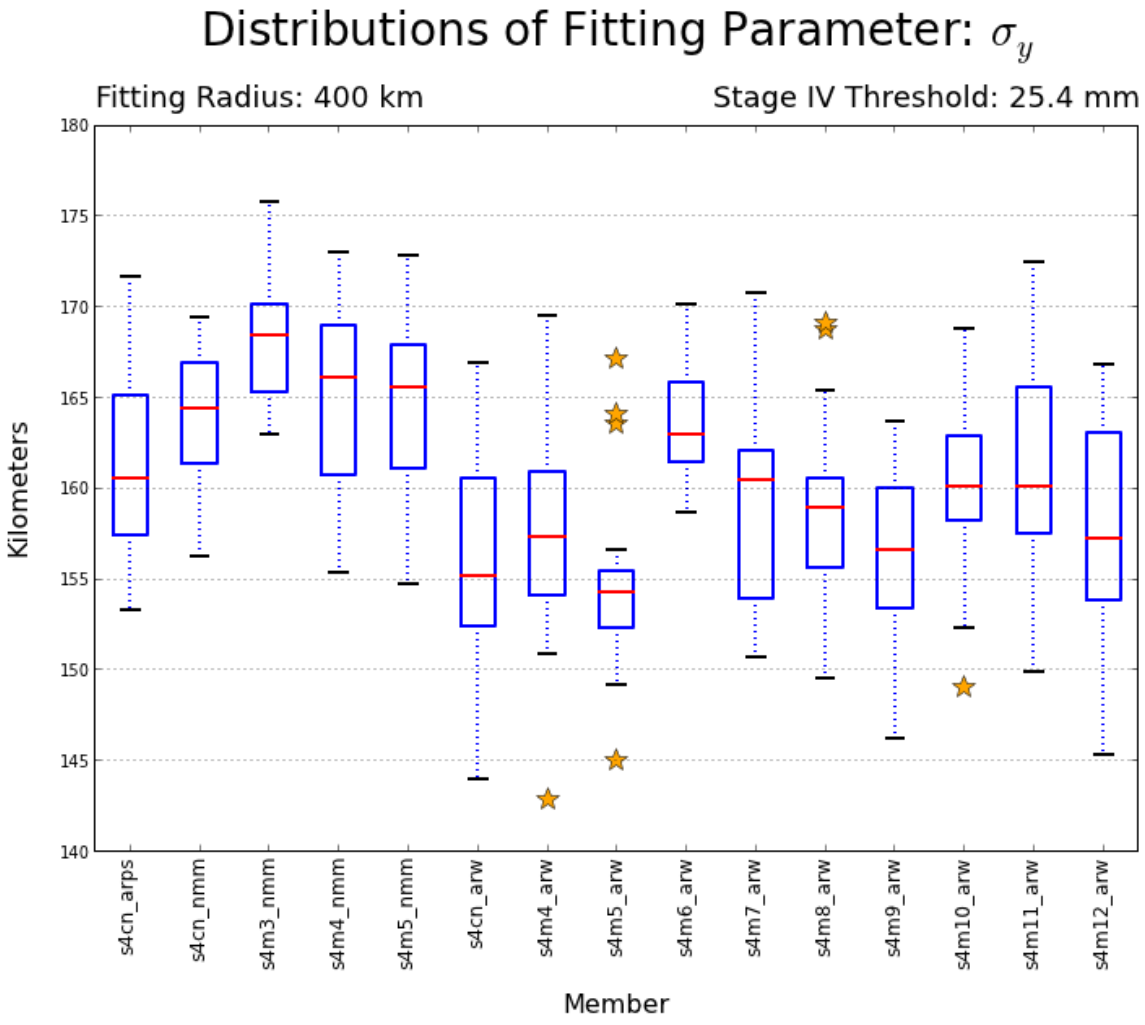

Figure 4.2: The same as in Figure 4.1, except for fitting parameter σ_y.

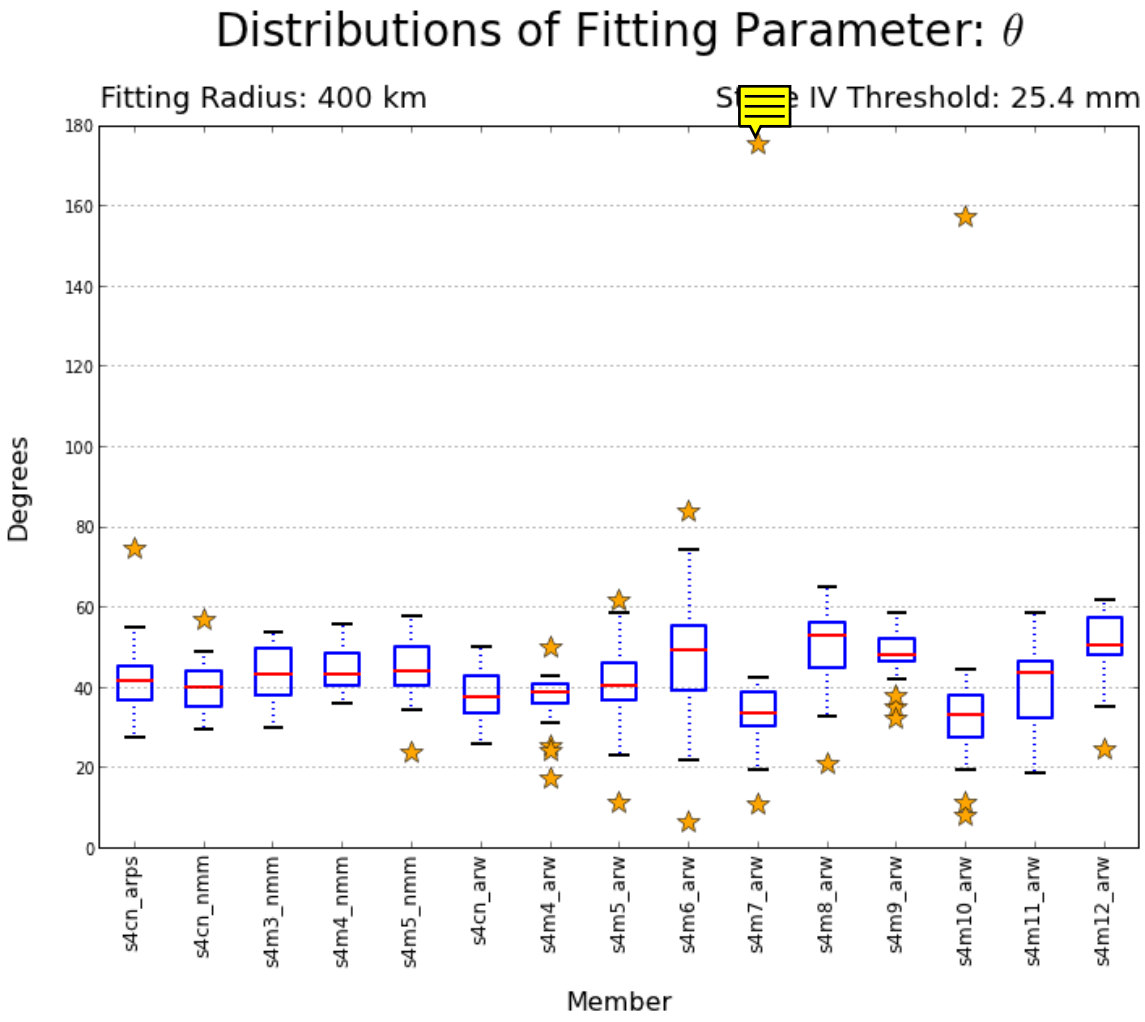

Figure 4.3: The same as in Figure 4.1, except for fitting parameter θ.

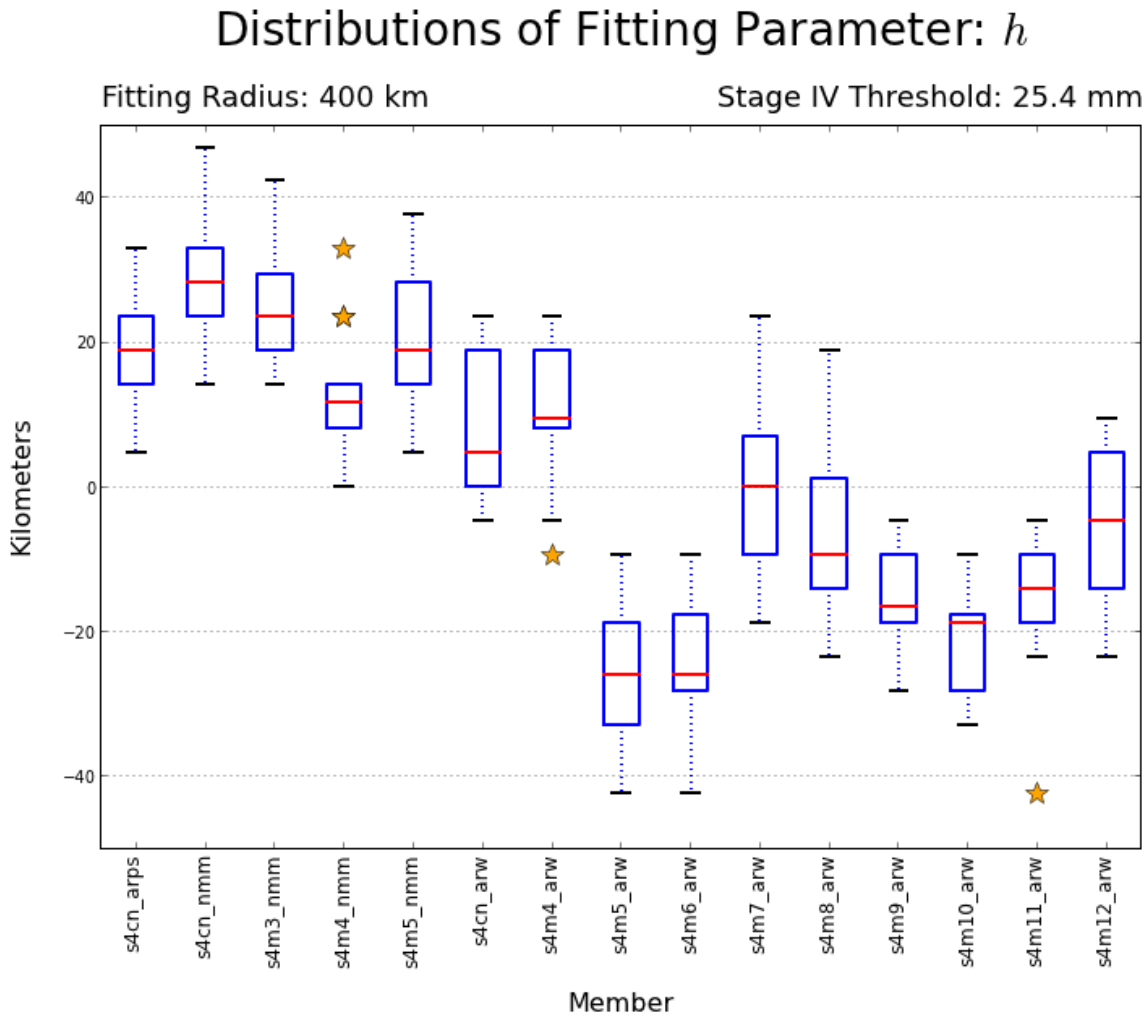

Figure 4.4: The same as in Figure 4.1, except for fitting parameter h.

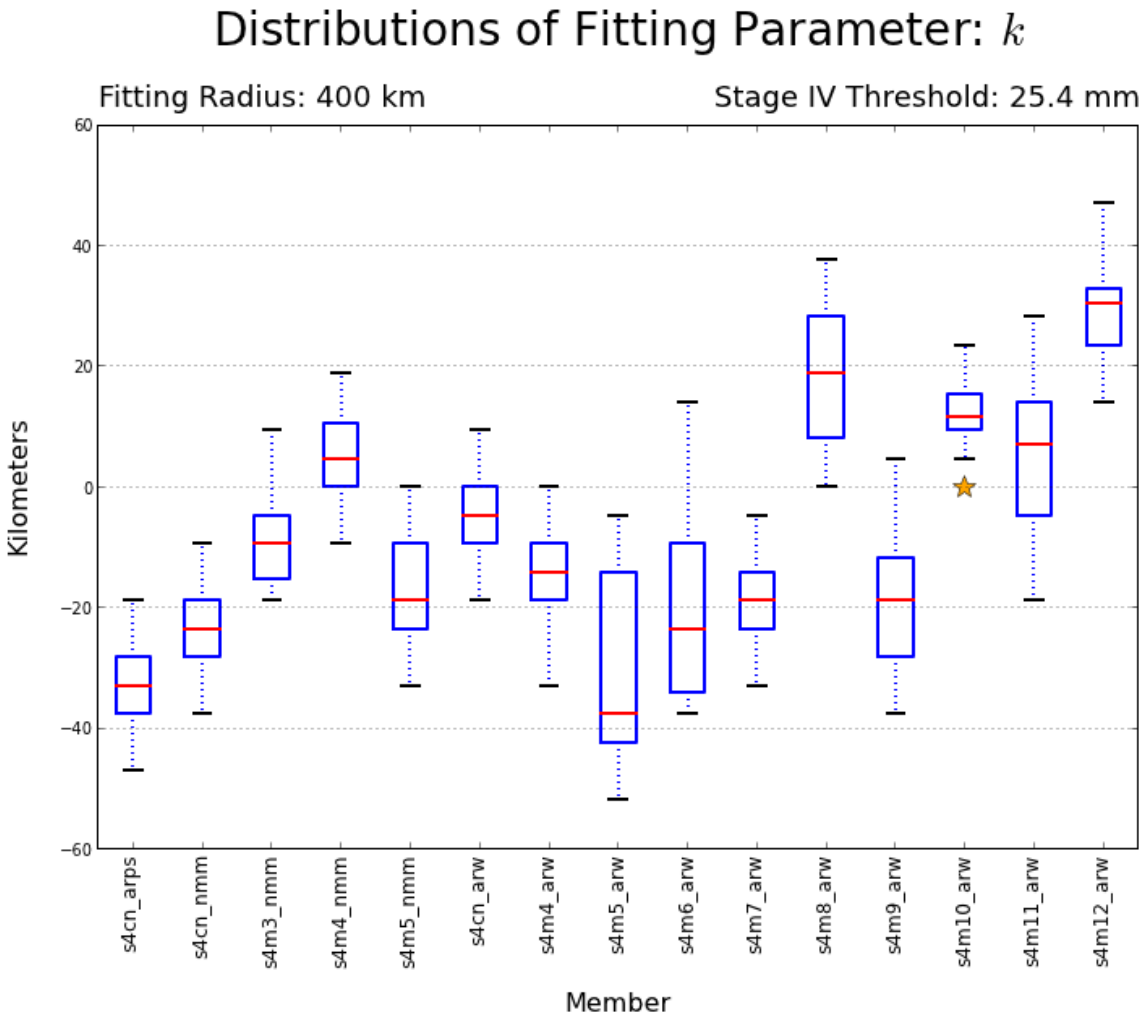

Figure 4.5: The same as in Figure 4.1, except for fitting parameter k.

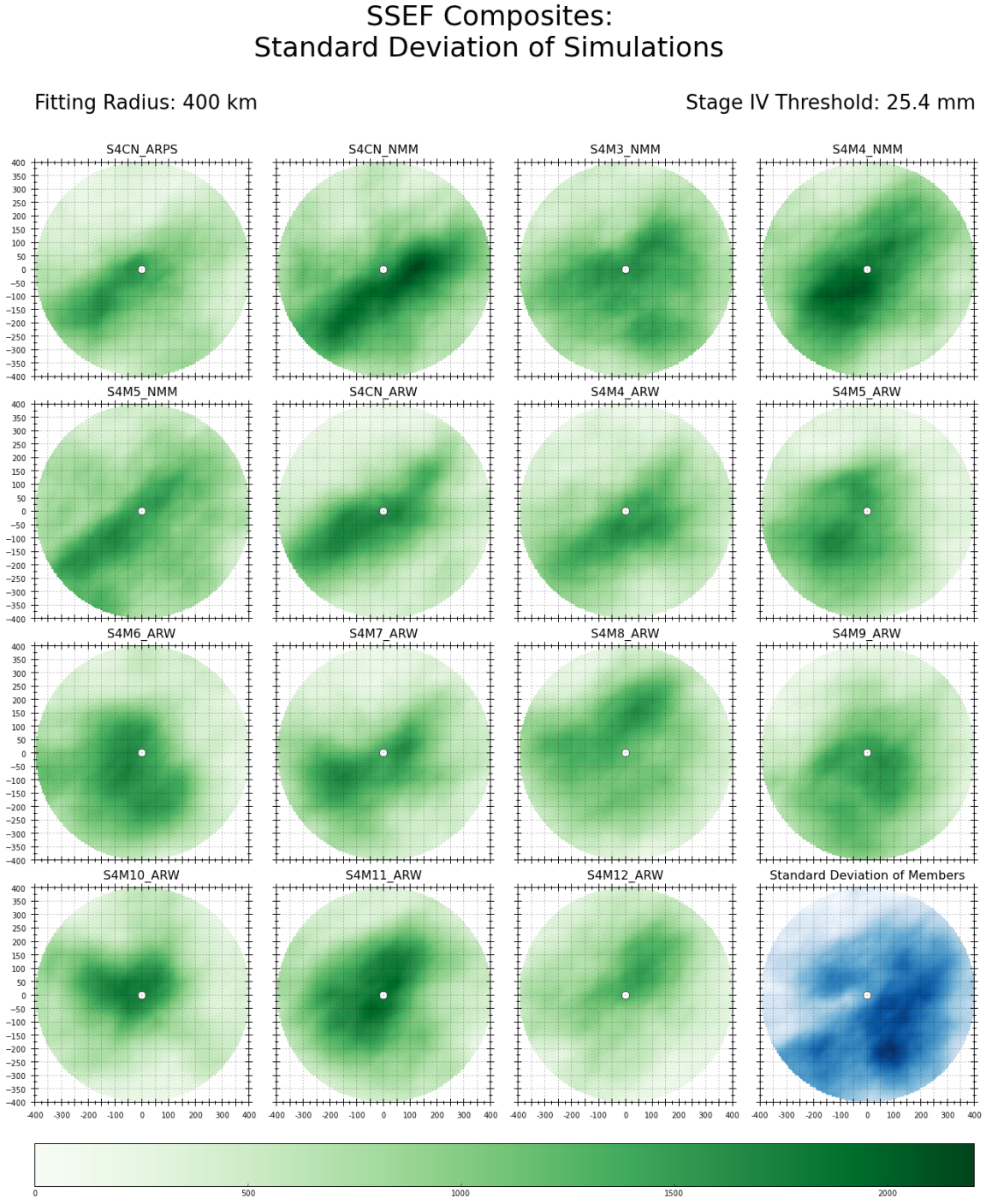

Figure 4.6: Plots of the standard deviation of the number of observations at each grid point, for each member of the SSEF at the 25.4 mm in 6 hr threshold. The standard deviation represents variability in the number of observational counts at each grid point between each of the twenty simulations. The lower right panel is a depiction of the standard deviation at each grid point for all members and all simulations. The color scale is different than all other panels to indicate the scale for this panel is different from the scale for all other members. A scale for the lower right panel is not shown as its purpose is to be qualitative instead of quantitative.

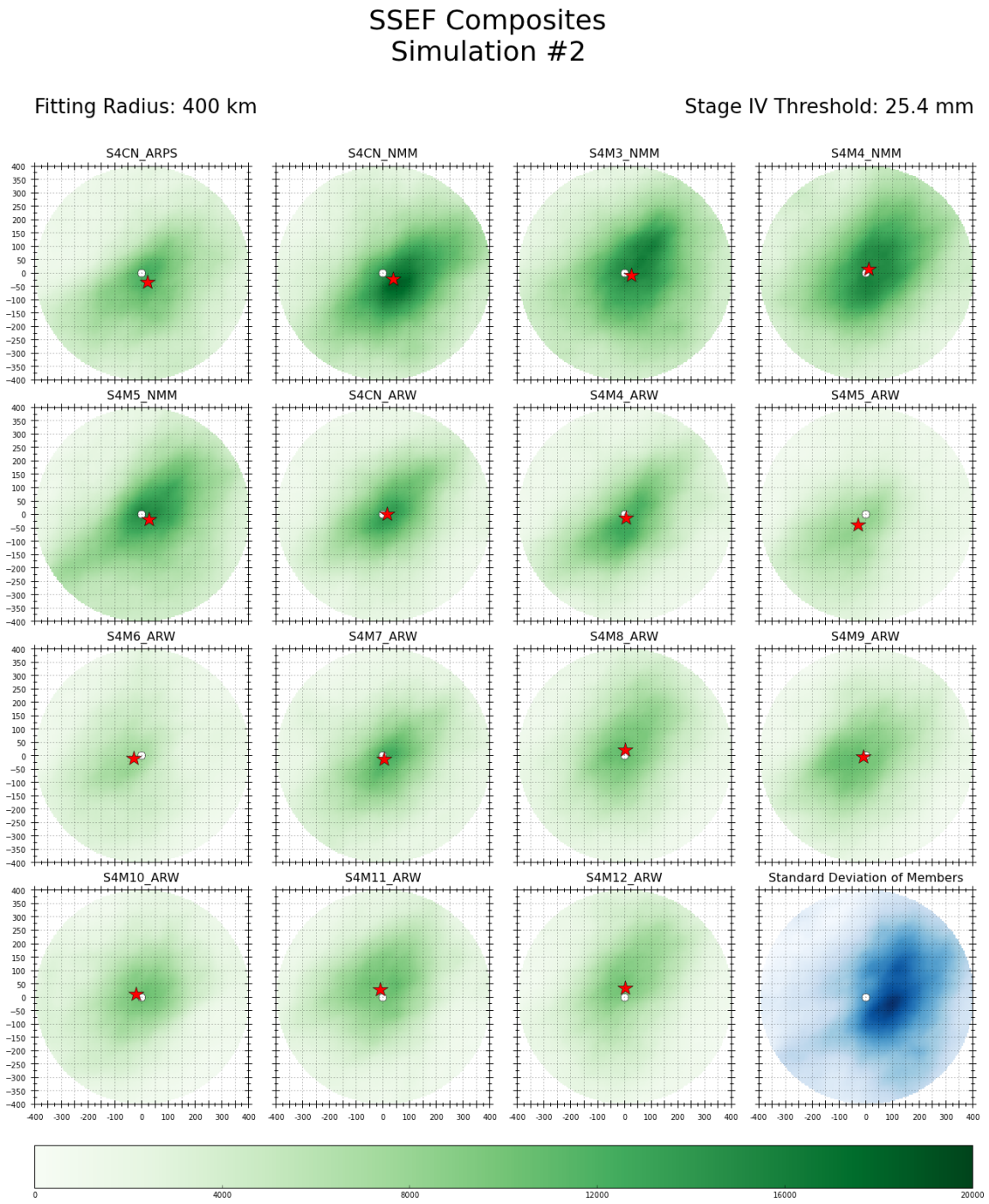

Figure 4.7: Plots of the two-dimensional composites at the 25.4 mm in 6 hr threshold for each member of the SSEF for a specific simulation. The lower right panel qualitatively depicts the standard deviation at each grid point for each of the two-dimensional composites shown in the other panels. As was the case in Figure 4.6, the color scale is different from the others to prevent comparison with the other panels.

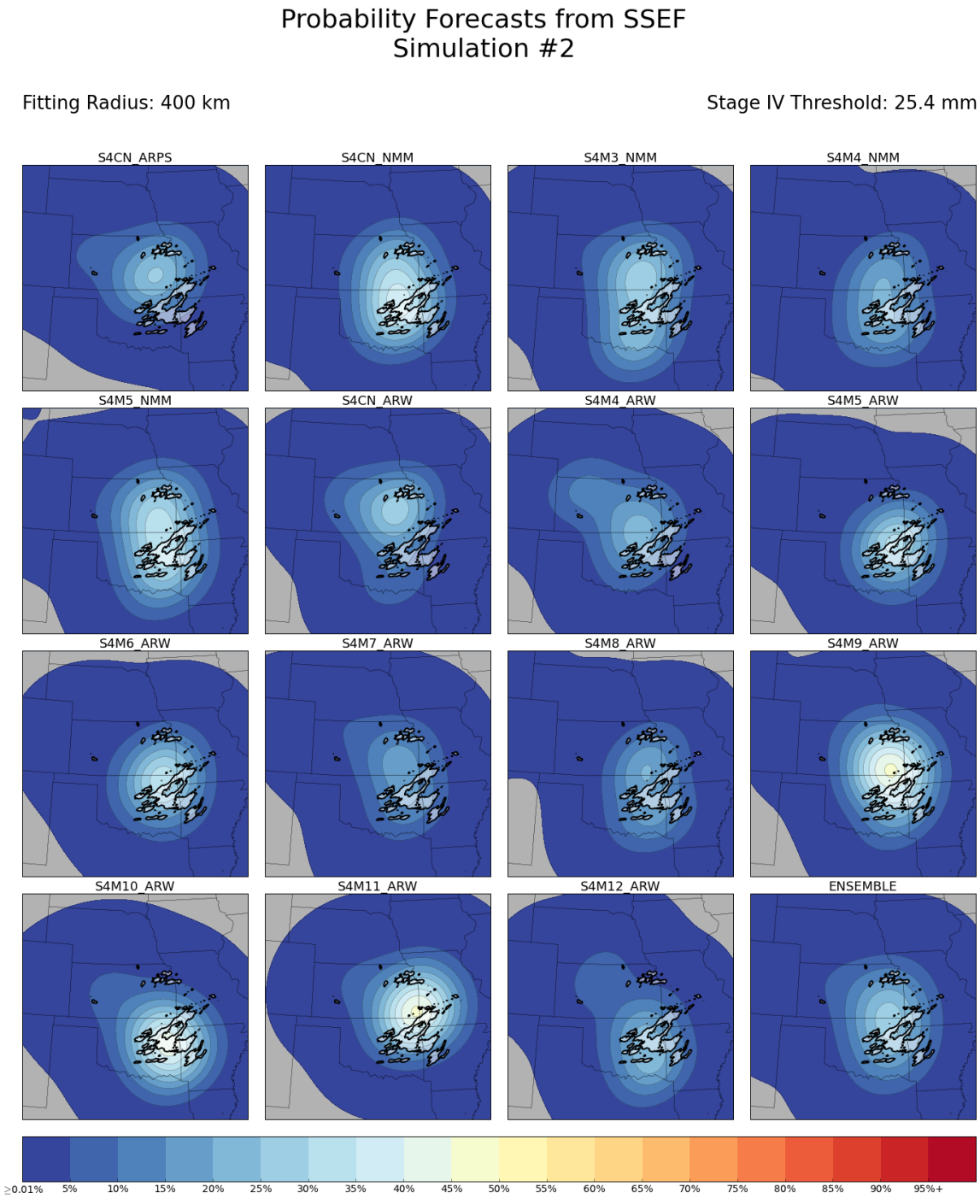

Figure 4.8: Example probabilistic forecasts of exceeding 25.4 mm in 6 hr from each SSEF member for the six hours ending 06 UTC 20 May 2010. The Stage IV QPE greater than 25.4 mm is contoured on top of the probability forecasts for each member. The lower right panel depicts the ensemble average probability.

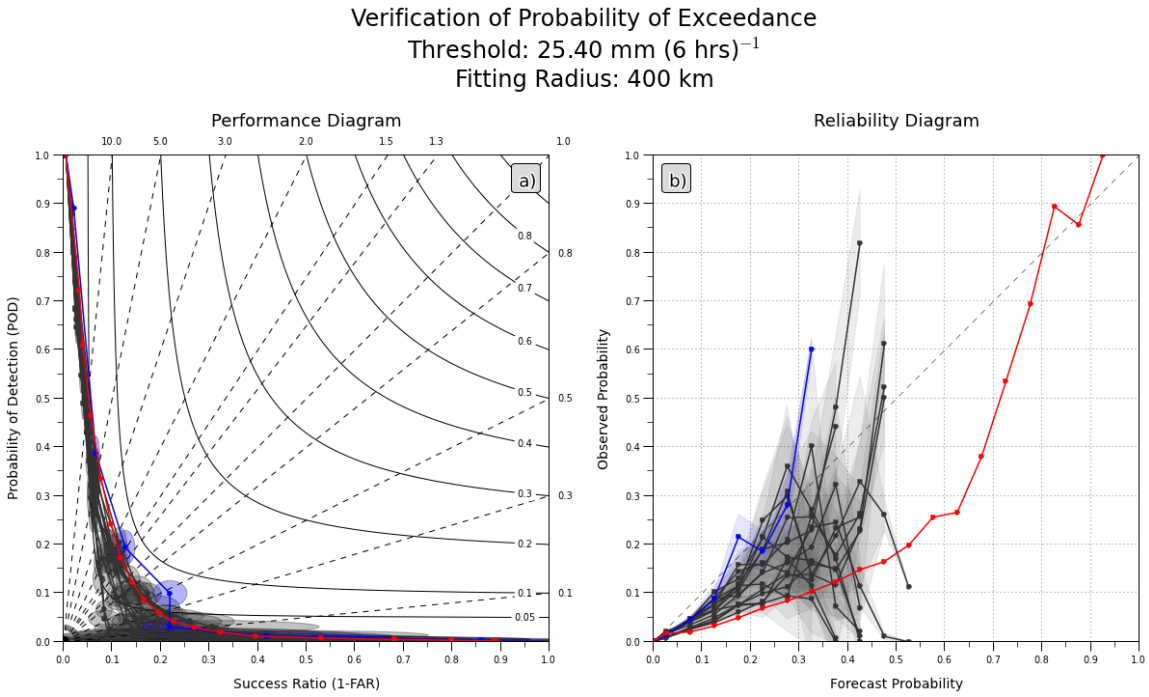

Figure 4.9: Performance Diagram (a) and reliability diagram (b) for each member of the SSEF for the 25.4 mm in 6 hr. The line of perfect reliability (diagonal; dashed) is also plotted on the reliability diagram. The mean values for each individual SSEF members' twenty simulations are shown in black, with standard errors in light gray. The ensemble average forecast verification is shown in blue. The red curve depicts the verification of the modified Hamill and Colucci porbabilistic forecasts.

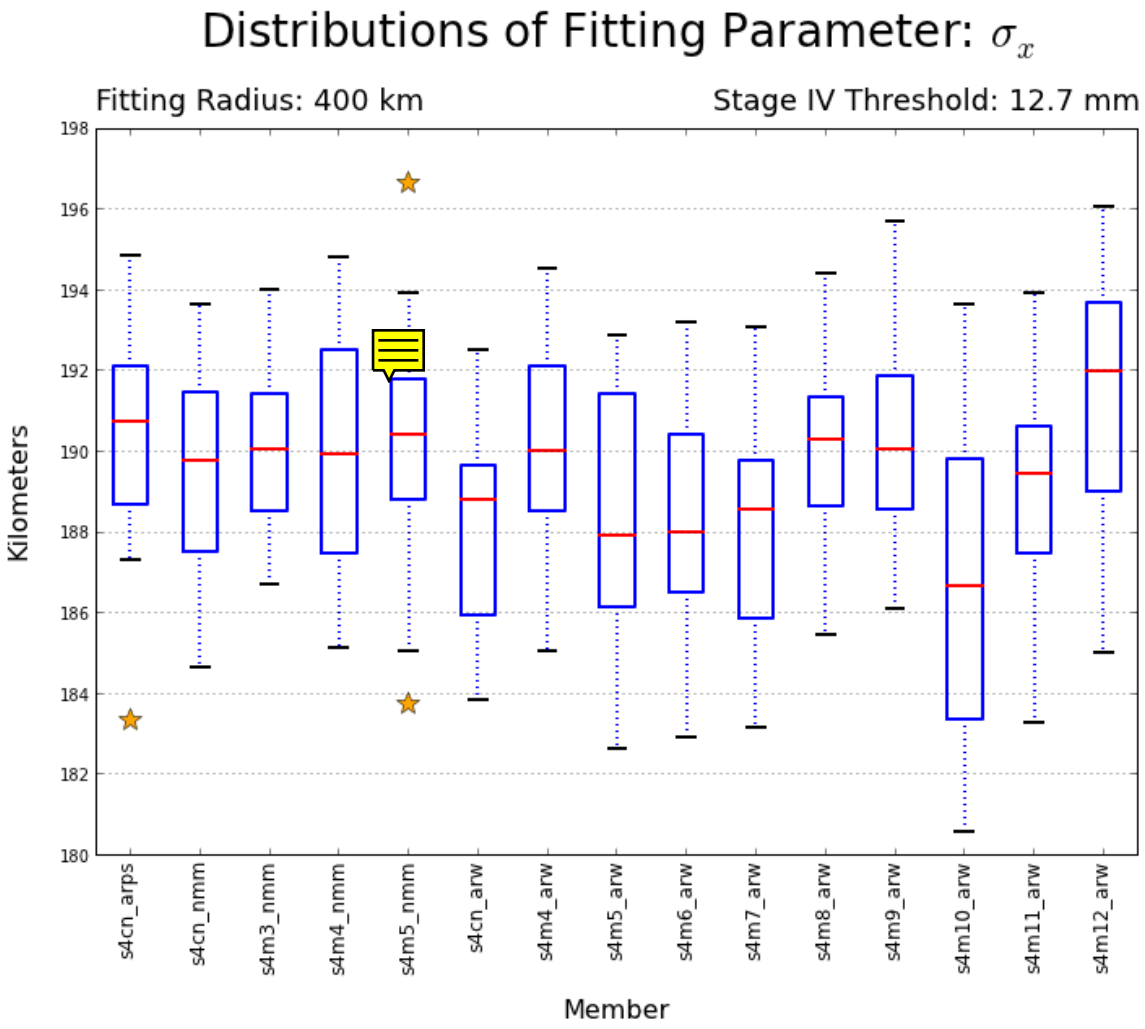

Figure 4.10: The same as in Figure 4.1, except for threshold 12.7 mm in 6 hr.

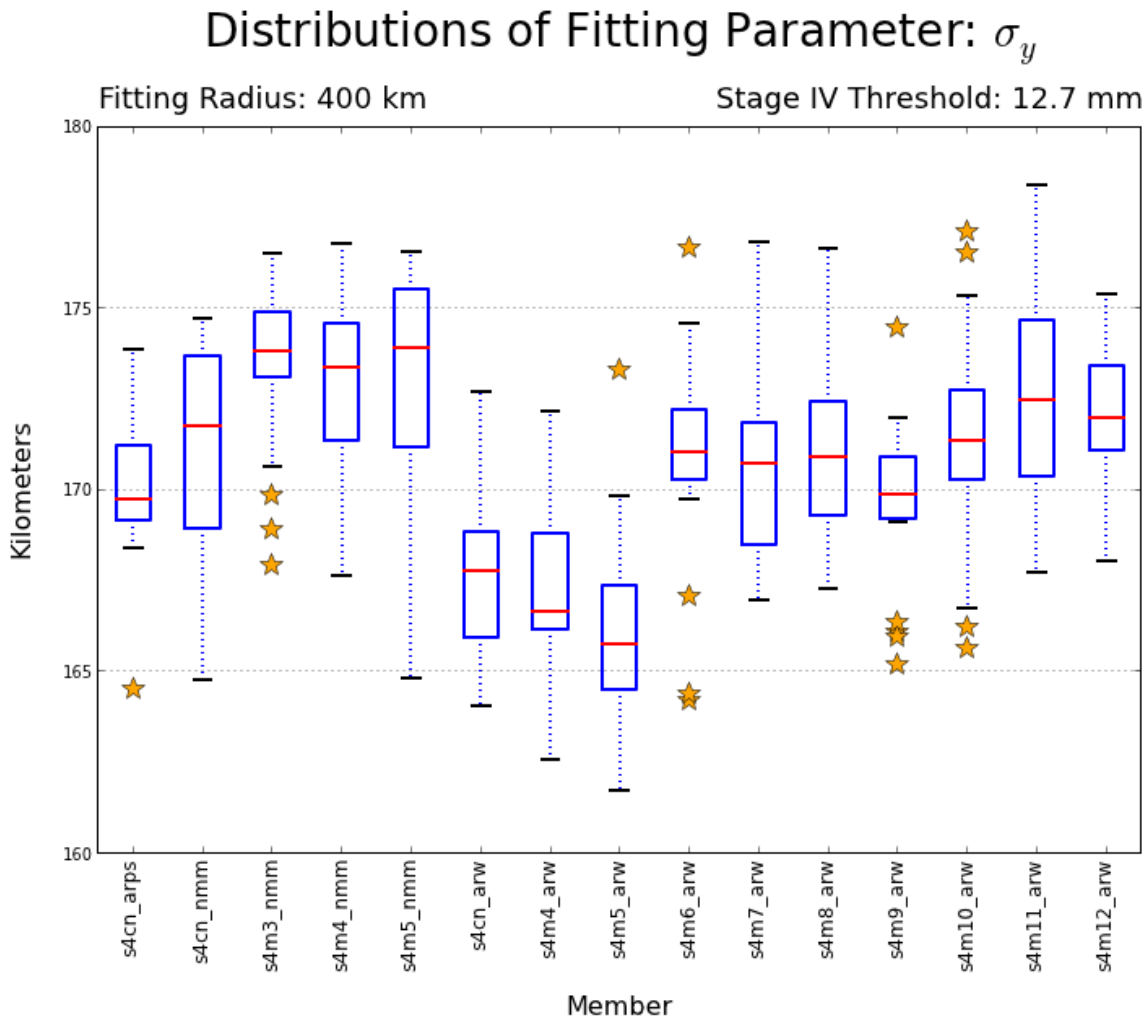

Figure 4.11: The same as in Figure 4.2, except for threshold 12.7 mm in 6 hr.

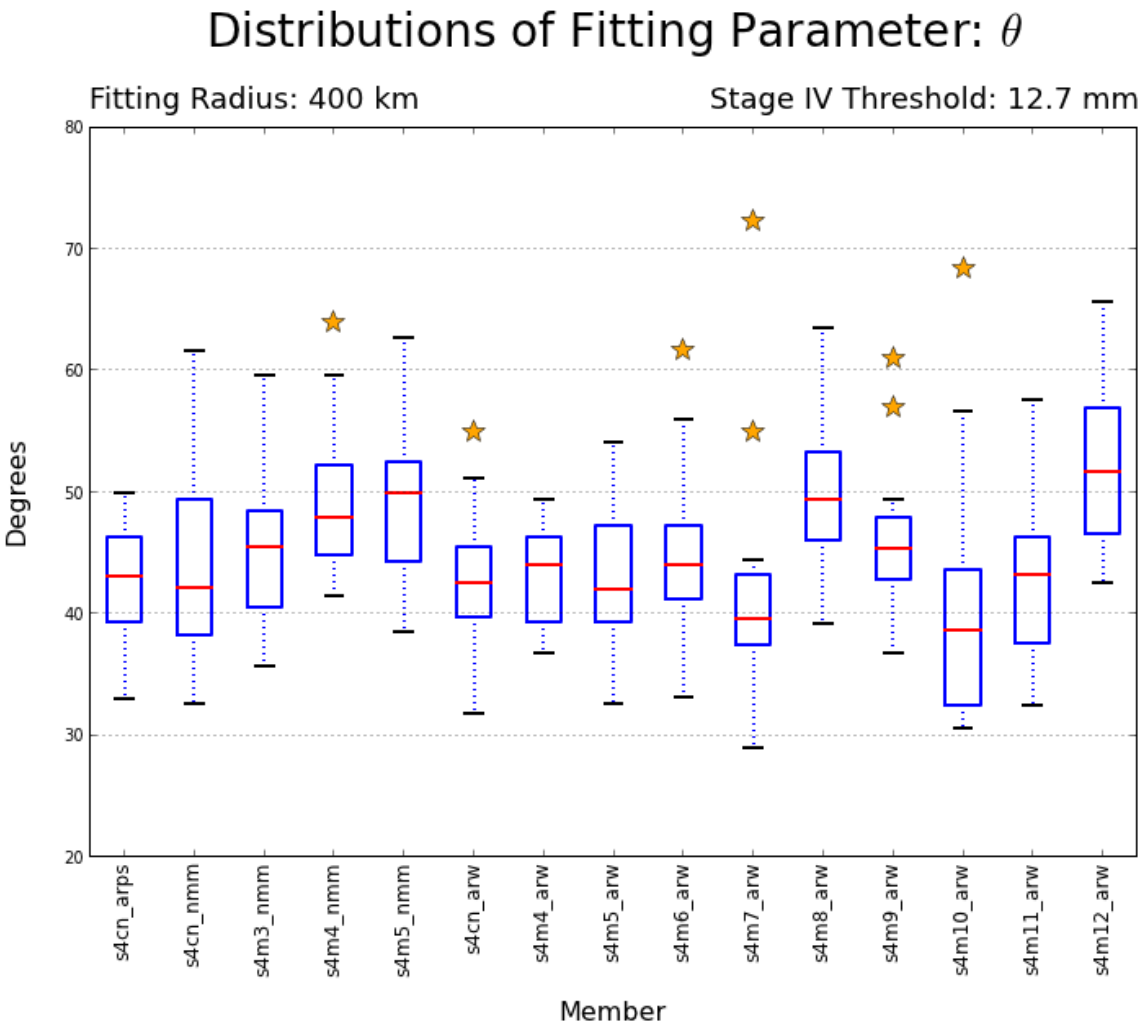

Figure 4.12: The same as in Figure 4.3, except for threshold 12.7 mm in 6 hr.

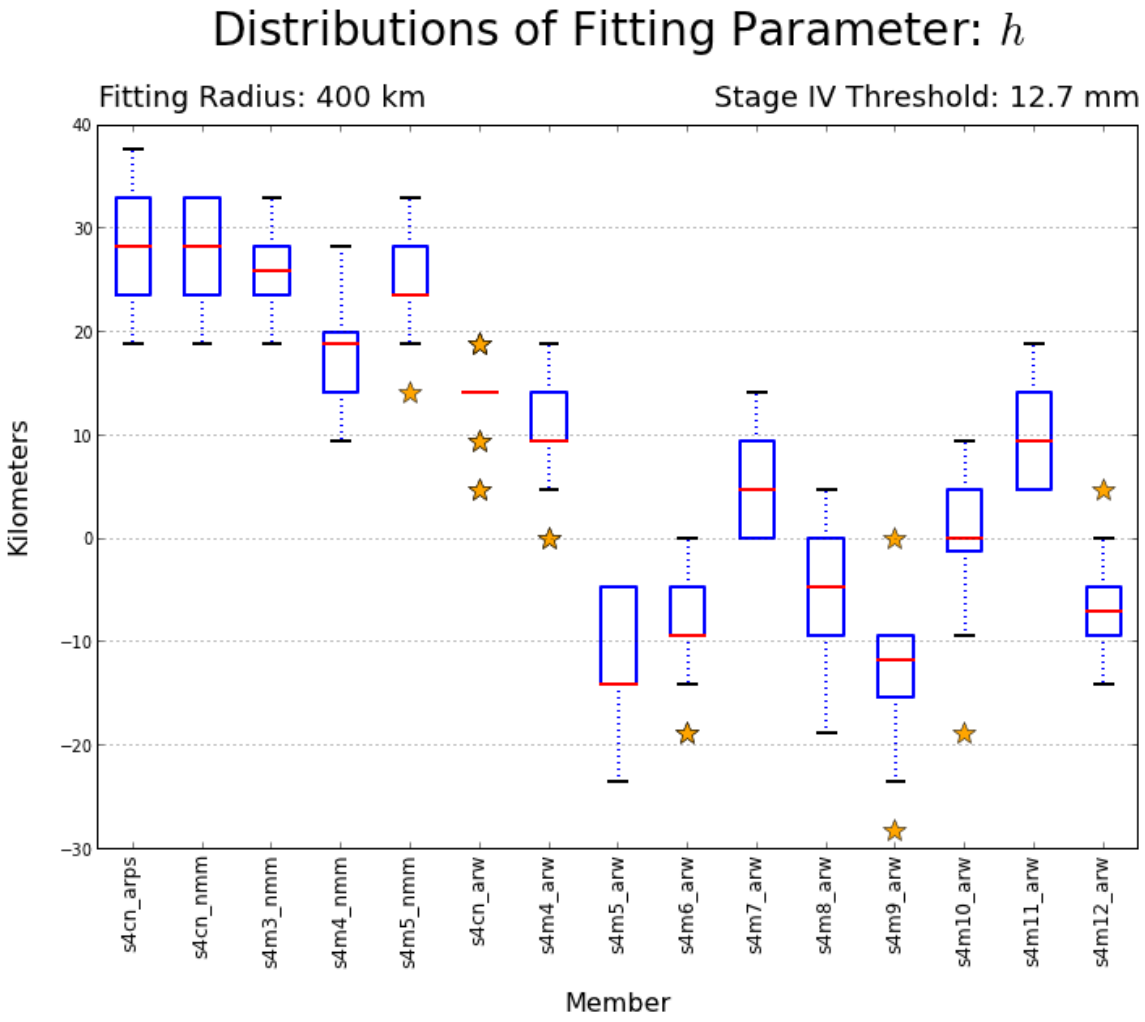

Figure 4.13: The same as in Figure 4.4, except for threshold 12.7 mm in 6 hr.

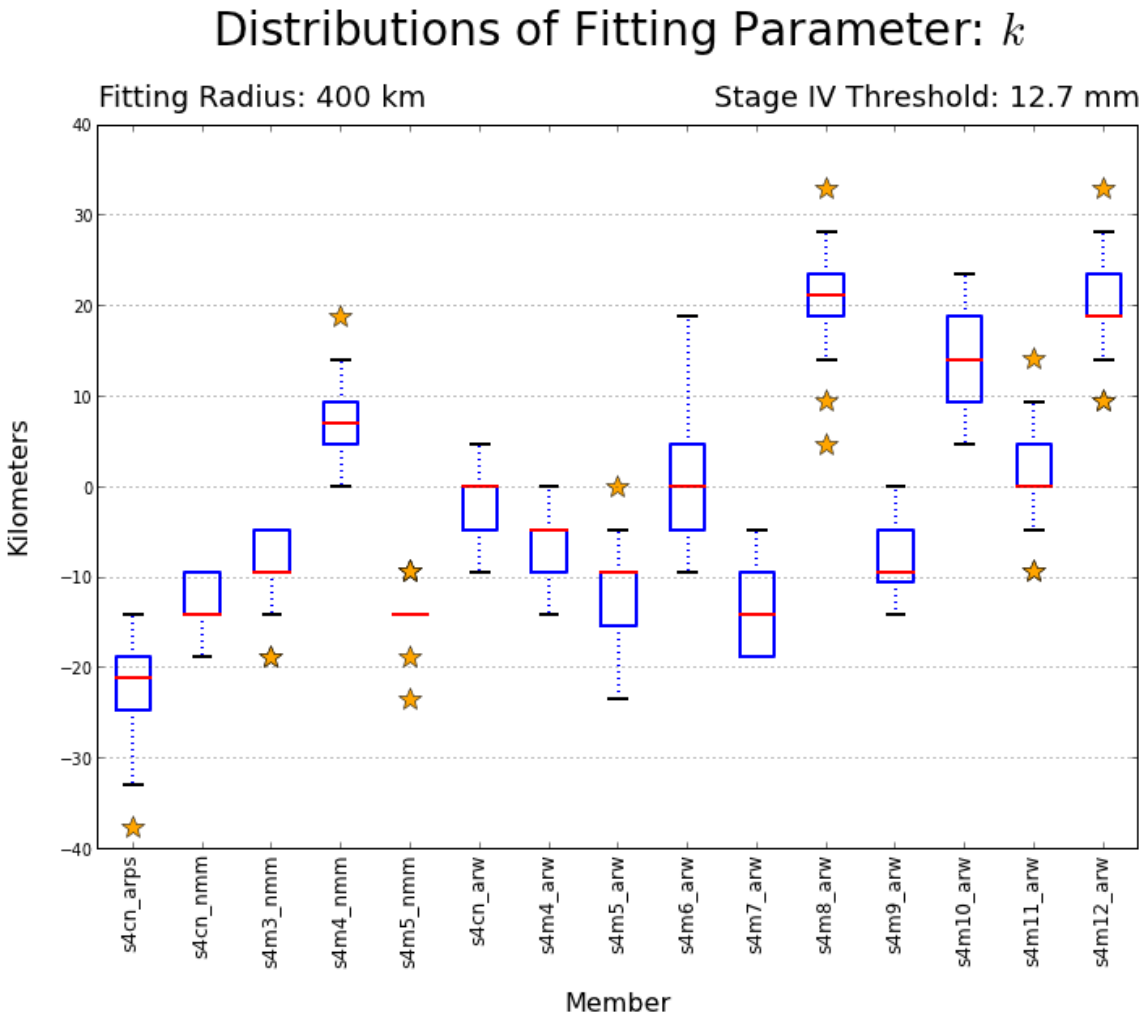

Figure 4.14: The same as in Figure 4.5, except for threshold 12.7 mm in 6 hr.

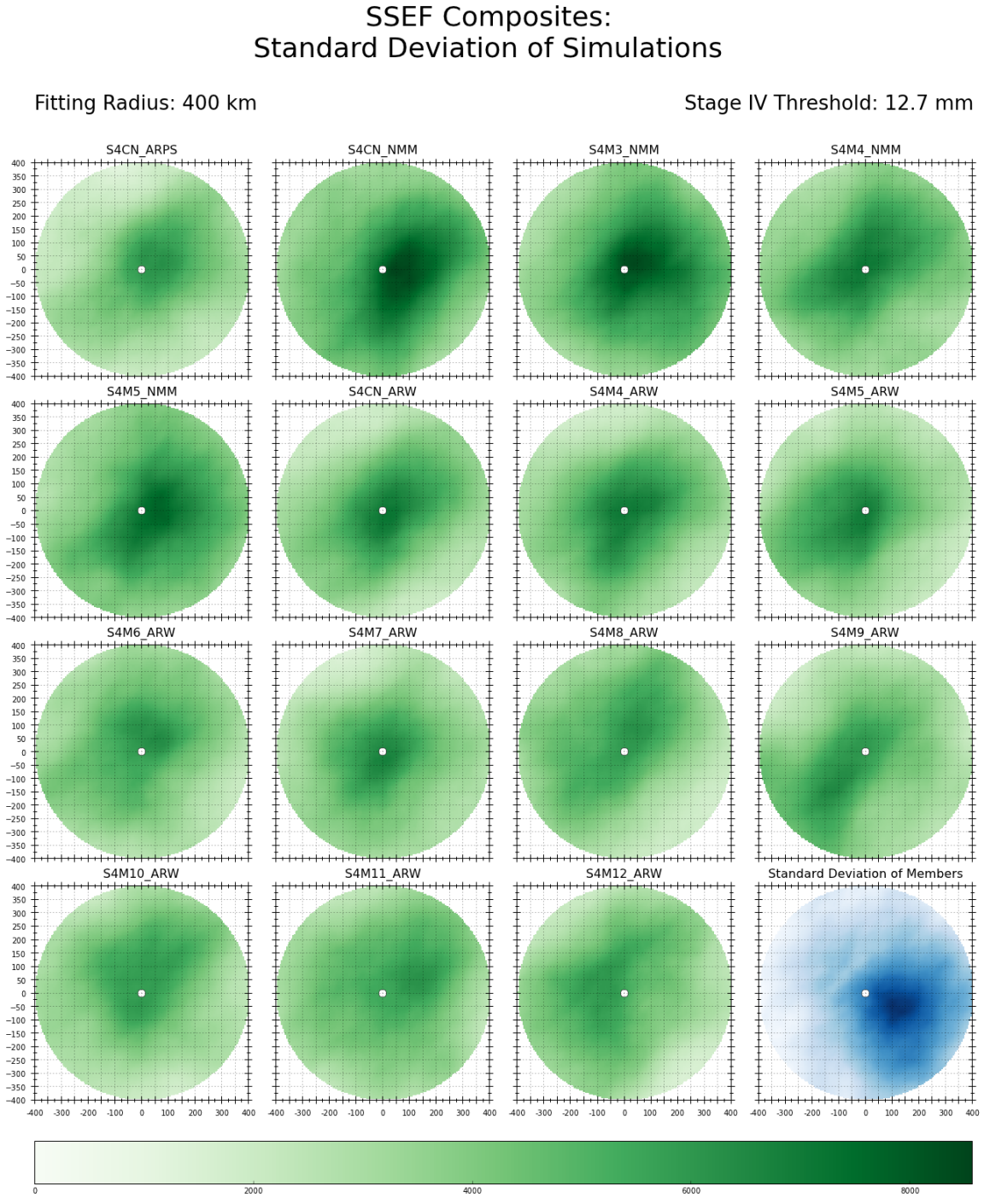

Figure 4.15: The same as in Figure 4.6, except for threshold 12.7 mm in 6 hr.

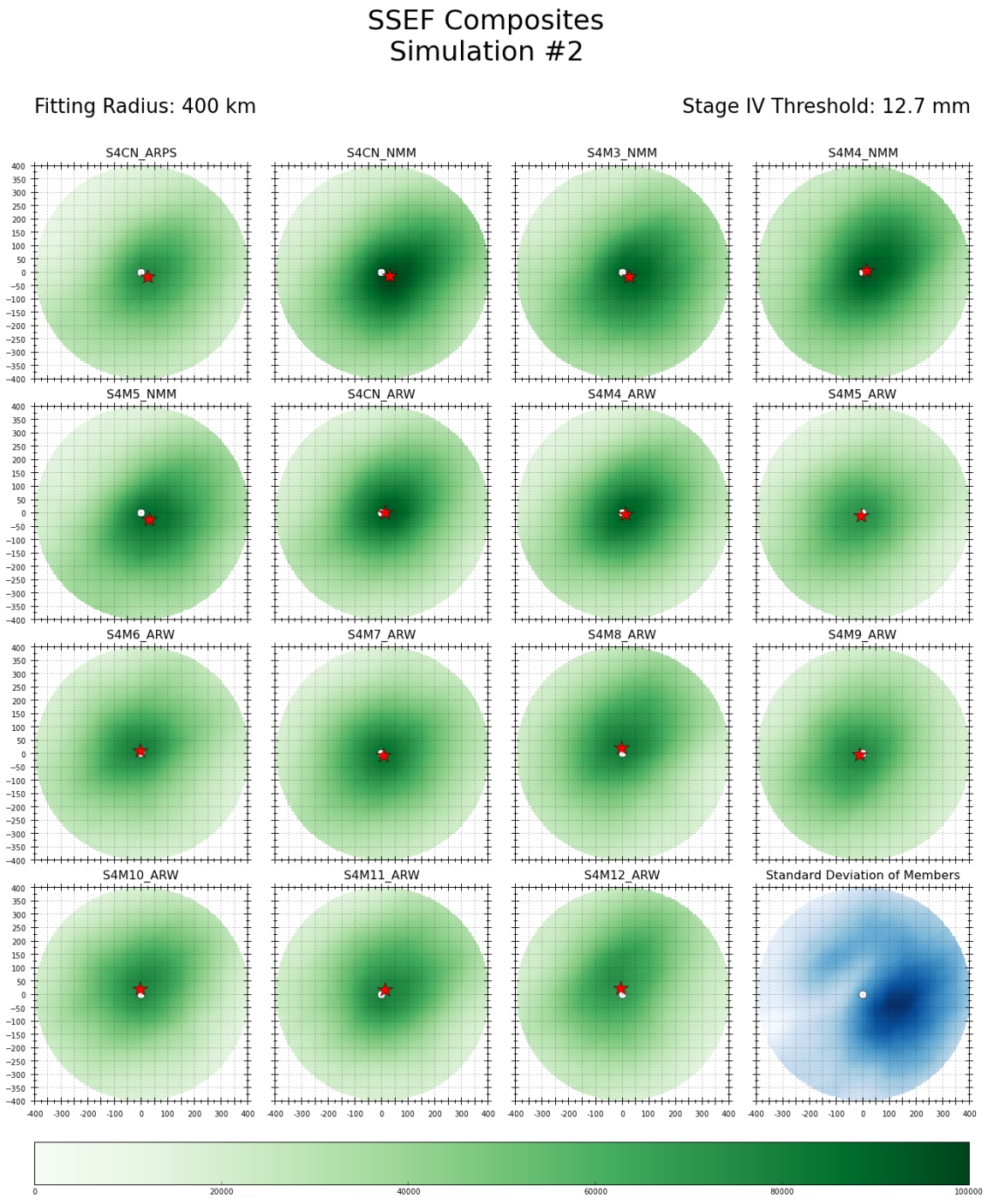

Figure 4.16: The same as in Figure 4.7, except for threshold 12.7 mm in 6 hr.

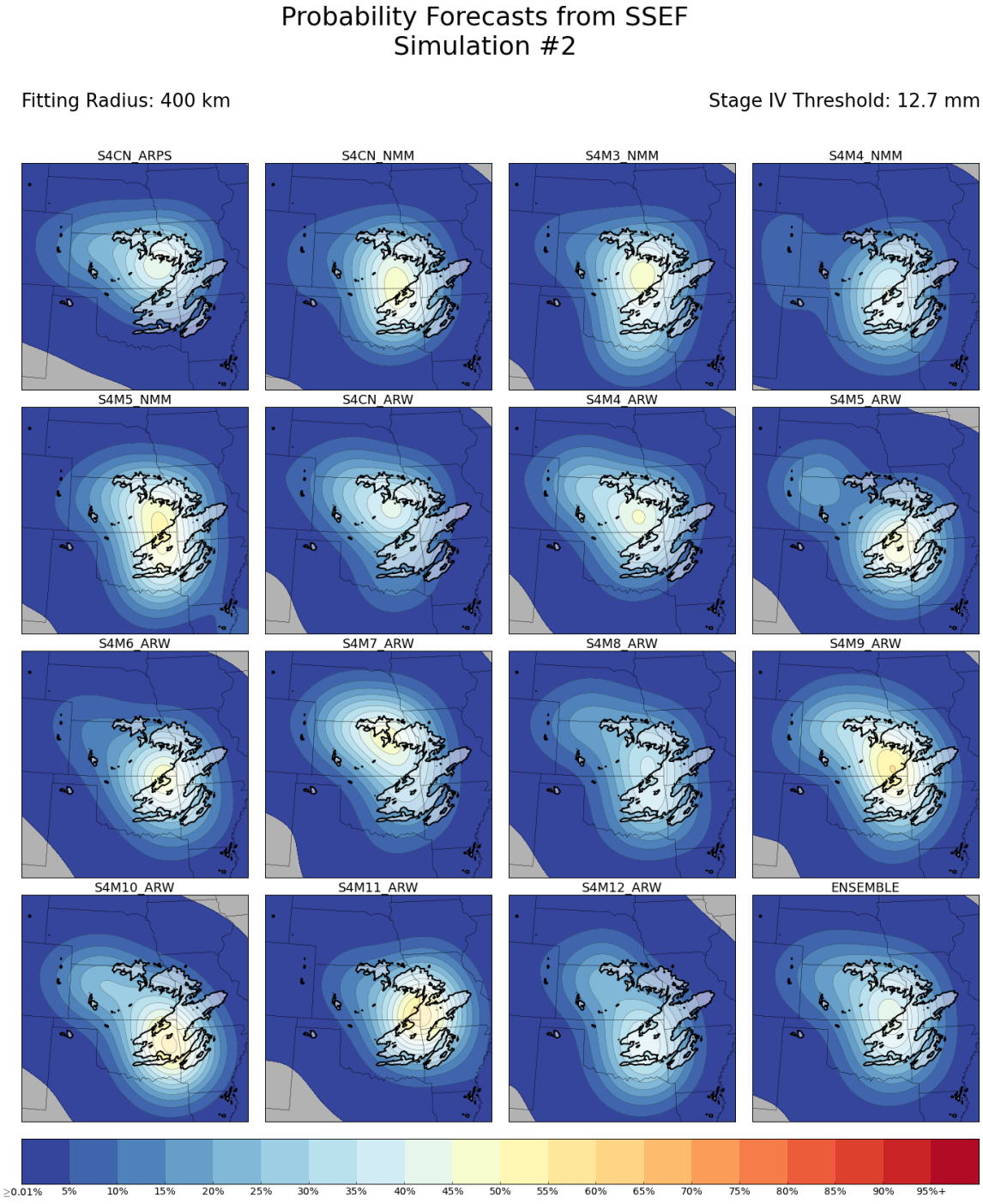

Figure 4.17: The same as in Figure 4.8, except for threshold 12.7 mm in 6 hr.

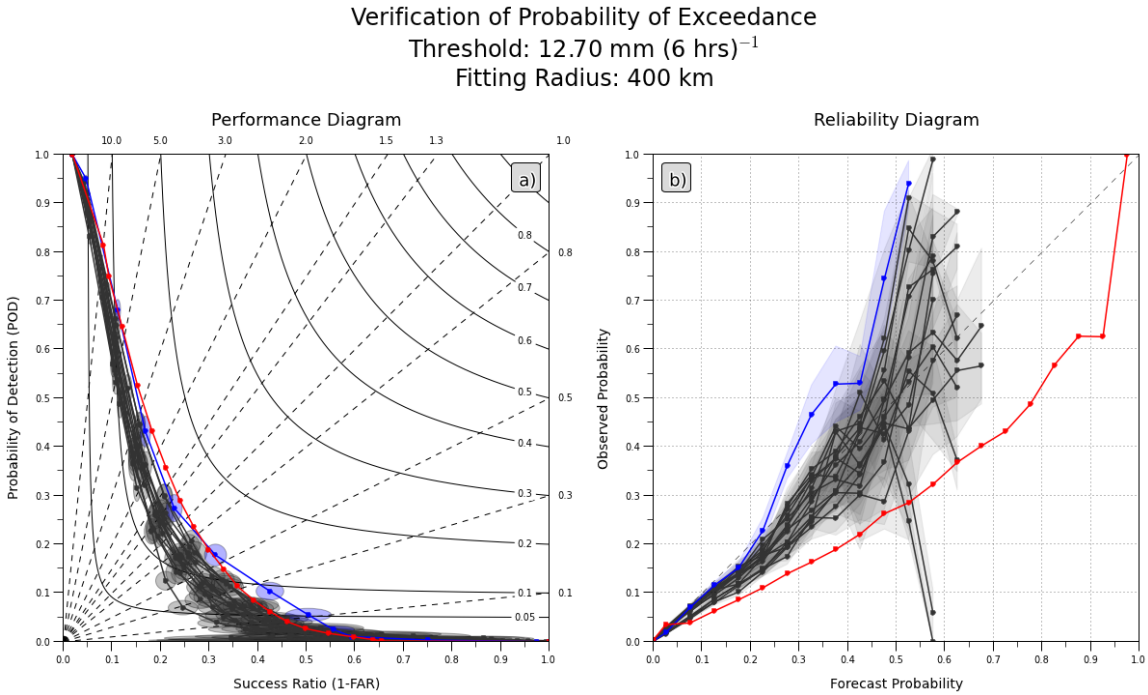

Figure 4.18: The same as in Figure 4.9, except for threshold 12.7 mm in 6 hr.

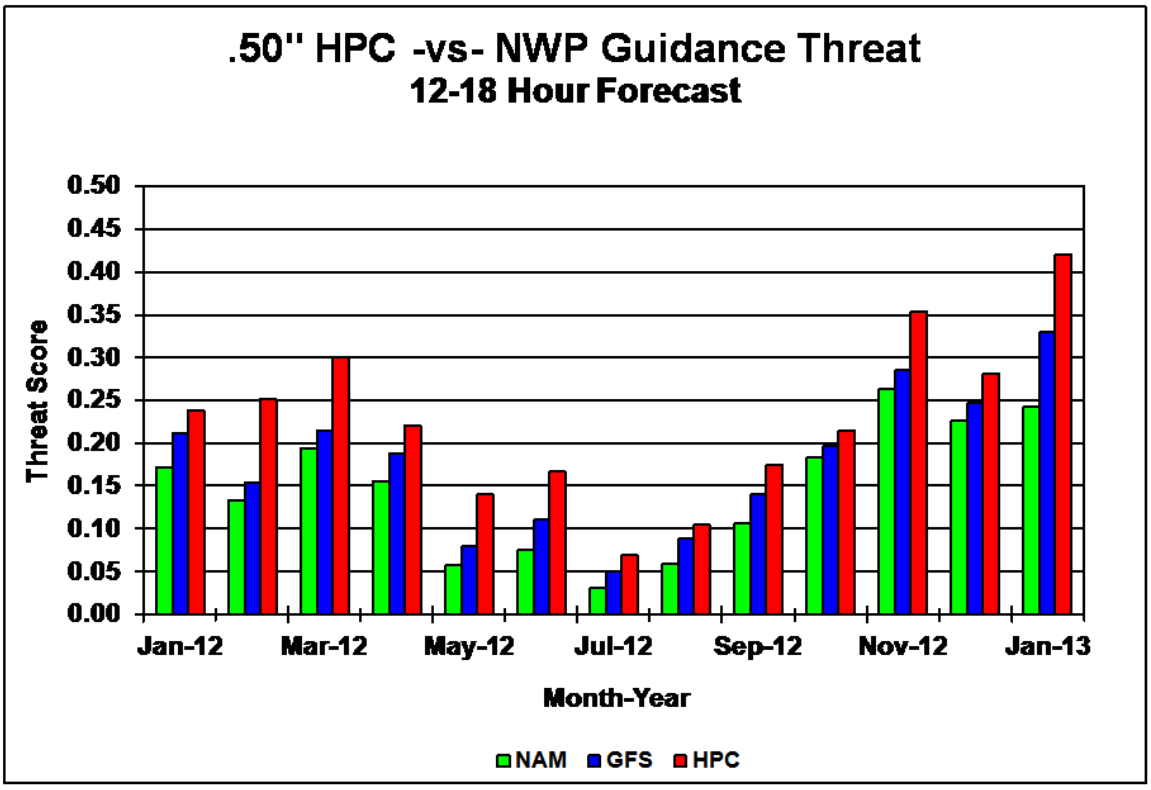

Figure 4.19: Quantitative precipitation forecast verification scores at the 12.7 mm in 6 hr threshold for the GFS and NAM numerical models and the HPC human forecasts. This is for forecasts with lead time of 12-18 hr.

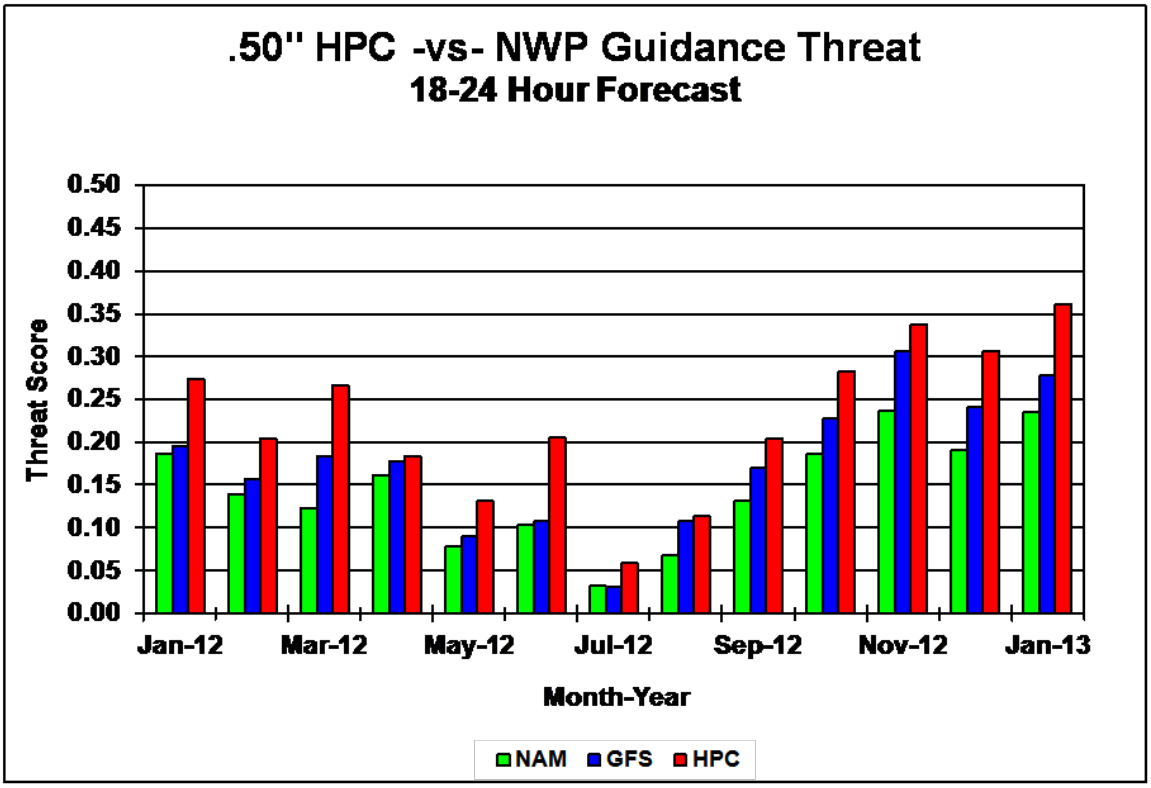

Figure 4.20: The same as in Figure 4.19, except for forecasts with lead time of 18-24 hr.

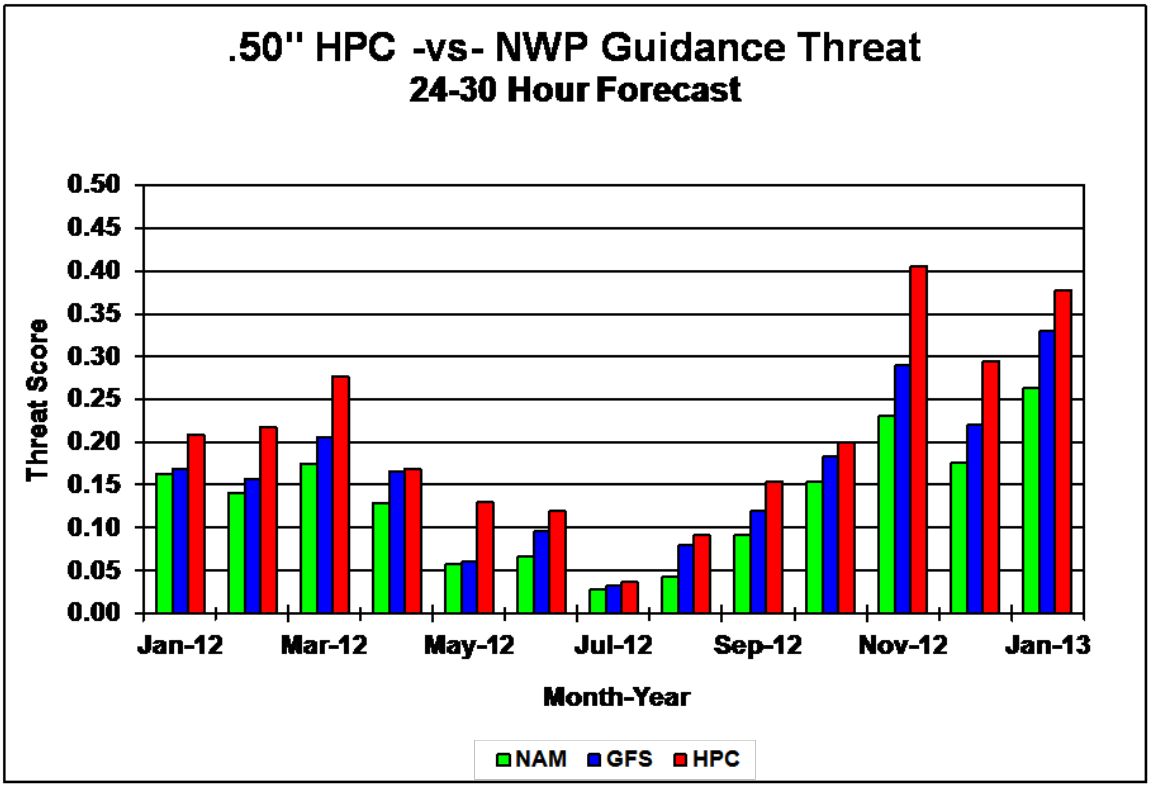

Figure 4.21: The same as in Figure 4.19, except for forecasts with lead time of 24-30 hr.

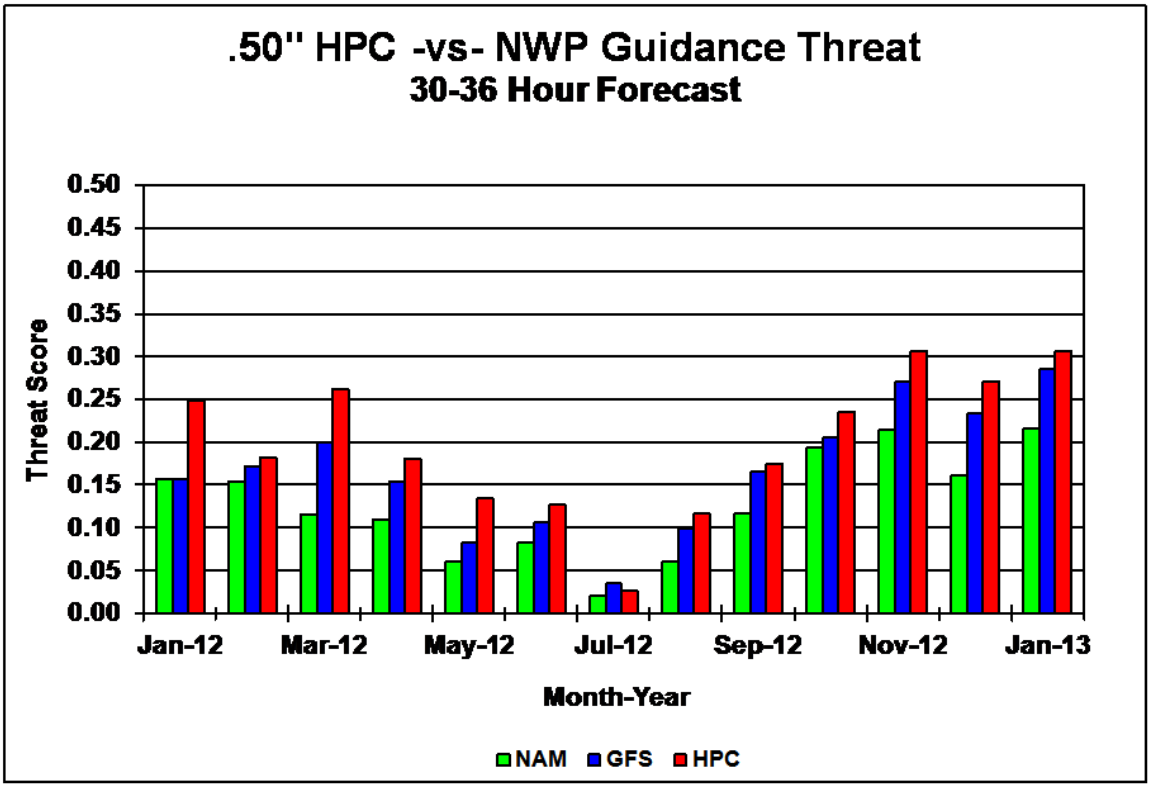

Figure 4.22: The same as in Figure 4.19, except for forecasts with lead time of 30-36 hr.

Chapter 5

Discussion

The motivation for this research was driven by the experiences with high resolution numerical models in the NOAA Hazardous Weather Testbed, in particular the NSSL-WRF, along with the needs of the "Warn-on-Forecast" (WoF) initiative. One aim of the WoF initiative is to transform the warning paradigm of rare convective events from one where RCE warnings are based almost entirely on observations to one where RCE warnings are based on short-term, high resolution numerical forecasts of RCEs. A key challenge for the WoF paradigm is to produce probabilistic guidance for the occurrence of RCEs that has a high degree of statistical reliability and resolution and is unambiguous for users to interpret. This is especially difficult since these phenomena will not be explicitly resolved in larger domain model configurations for many years to come (e.g., explicit prediction of tornadoes will require grid spacing on the order of a few tens of meters).

One possibility for overcoming this problem is to identify "extreme" model-generated features that have strong correlations with observed severe convective phenomena, and then use the former as surrogates for the severe phenomena in question. This "surrogate-severe" (SS) approach is fundamentally different from traditional applications of numerical weather prediction for severe weather because it is phenomenon based. In particular, it relies on identification of explicit convective phenomena rather than environmental conditions that might support such phenomena. Sobash et al. (2011) established the viability of this approach using several different SS diagnostic quantities. Among the quantities they examined, model-generated updraft helicity appeared to show the strongest correla-

tion with observed reports of severe weather. [Updraft helicity is a measure of mid-level rotation in model-predicted updrafts and subjective assessments suggest that it is a useful surrogate for supercell thunderstorms (Kain et al. 2010), even when these storms are only crudely represented on the WRF model's native grid (Kain et al. 2008).] Subjective assessments in the HWT convinced participants that SS quantities from high resolution models had the potential to offer guidance to forecasters as to the vicinity (in time and space) of RCE occurrence, but not necessarily the exact location.

The fact that subjective assessments of high resolution numerical model forecasts suggest that forecasts of RCEs are in the vicinity of observations of RCEs, but are not necessarily collocated, highlights the need to communicate the uncertainty in the forecast. One way to do this is to utilize an ensemble of high resolution numerical models to quantify the spatial uncertainty in the location of the forecast RCEs. Unfortunately, the infrequent nature of rare events makes it unlikely that two separate high-resolution model forecasts would place extreme model-generated convective storm phenomena at the same grid point, even for generally similar mesoscale forecasts. Thus, ensemble generated probabilities of RCE occurrence at a given grid point are typically extremely small. This is consistent with the limited predictability on the convective scale and the associated low climatological frequency of rare events, which makes it difficult to convey statistically meaningful severe weather threats to the user community (Murphy 1991).

Informal conversations with operational meteorologists in the HWT suggest that both forecasters and users of hazardous weather information may not respond appropriately to the very small probability values that result from creating ensemble probabilities of RCE occurrence on a fine grid, as is the cause with storm scale ensemble forecast systems. One potential remedy to this problem was offered by Sobash et al. (2011). Instead of using an ensemble to generate a probabilistic forecast, their method utilized a a single deterministic forecast and applied a "neighborhood"-based approach that is rooted in the concepts of Theis et al. (2005) and Brooks et al. (1998).

This neighborhood approach consists of two steps. The first step involves taking binary grid point forecasts of occurrence of specific events and expanding their spatial extent by converting all grid points within a specified "neighborhood" into forecasts of the given event's occurrence. For example, consider a case in which a single grid point from a high resolution numerical model (grid spacing of 4 km) is forecast to have a phenomenon occur. Furthermore consider that a neighborhood of 40 km is specified The first step of the neighborhood approach takes all grid points within the specified neighborhood of the forecast grid point, 40 km radius in this example, and converts those grid points into forecasts of the phenomenon's occurrence. This effectively increases the area for which the phenomenon is forecast.

The second step involves using kernel density estimation to convert the neighborhood forecasts into forecasts of probability of occurrence. Sobash et al. (2011) employed a two-dimensional isotropic Gaussian kernel in their study, but had no method of optimally selecting the appropriate bandwidth. Their solution was to evaluate multiple choices of bandwidth and select the one that verified the best.

One negative to the neighborhood approach put forth by Sobash et al. (2011) is that by using a neighborhood, the specificity offered by high-resolution numerical models is diminished. No longer is a forecaster examining the probability of a phenomenon occurring on a given grid point, instead the forecaster is examining the probability of a phenomenon occurring within the defined neighborhood. However, given the limited predictability on the convective grid scale and the comparatively low probability values at the grid point, use of a neighborhood is typically considered an acceptable method to identify a region of enhanced threat.

Furthermore, the method put forth by Sobash et al. (2011) relies on accurate observations of the phenomenon being predicted to assess the quality of the resulting probabilistic forecasts. This poses significant limitations due to the lack of quality observations of these phenomena. As previously noted, numerical guidance of severe thunderstorms has

improved in recent years with the advent of convection-allowing models and high temporal resolution storm-attribute parameters (e.g., updraft-helicity, downdraft intensity, graupel loading, etc; Kain et al. 2010). Unfortunately, however, corresponding observational datasets with spatial and temporal coherence comparable to the model data are not available.

Quantitative precipitation forecasts, on the other hand, pose challenges to operational forecasters similar to those posed by severe thunderstorm forecasts. One important difference is that quantitative precipitation forecasts have comparatively robust verification datasets. Thus, one approach to improving the quality of numerical models' probabilistic forecasts of RCEs is to develop and refine techniques of predicting and calibrating extreme precipitation "events". After these enhancements have been fully evaluated using extreme precipitation events the refined methods can be applied to the original severe thunderstorm prediction problem.

This study begins this process. It is a proof-of-concept for using numerical forecasts of explicit phenomenon-based RCE to create objectively calibrated probabilistic forecasts. This builds on the work of Sobash et al. (2011) by using grid point forecasts of heavy precipitation events (defined to be either 12.7 mm or 25.4 mm in 6 hr) and the corresponding verification datasets. Furthermore, this is done without the use of a neighborhood. By forgoing the use of a neighborhood, forecasters can utilize the higher specificity offered by high resolution numerical models as compared to coarser resolution models such as the NAM. The objectively calibrated probabilistic forecasts are achieved by first computing a two-dimensional frequency distribution of observed precipitation events relative to forecasts of the same events. Next, these two-dimensional composites are used to determine the necessary parameters of an analytical function that can be used in the kernel density estimation step of Sobash et al. (2011). Assessing the utility of such an approach with two very different high resolution datasets — 1) 48-months of high resolution output from the real-time NSSL-WRF model; and 2) the individual member forecasts for the 119 time periods

from the 2010 and 2011 CAPS SSEF — generally indicates this technique has the potential to produce skillful probabilistic forecasts derived from a single numerical forecast.

During the 2011 NOAA HWT SFE, NSSL-WRF probabilistic forecasts of precipitation exceeding 25.4 mm in 6 hr, utilizing the approach put forth in this study, were subjectively evaluated During this subjective assessment, operational forecasters expressed concern about the resulting probability field, particularly the smooth appearance and low amplitude. Although a valid concern, the character of the probability fields is inherently linked to the underlying numerical model's ability to accurately predict the exact location of "events". When a model's two-dimensional histogram has a large area of relatively uniform observed event occurrence, indicative of relatively large model spatial uncertainty, the fitted analytic function generates broad, low-amplitude probabilities. Conversely, when the higher observed event occurrence in the two-dimensional histogram is more concentrated, indicating a reduction in model spatial uncertainty, the analytic function produces sharper, higher-amplitude probabilities. Simply stated, this technique objectively quantifies the spatial uncertainty associated with the deterministic forecasting skill of the modeling system. Thus, concerns about overly smooth, low-amplitude probability fields are directly related to the deterministic model's ability to accurately predict the location of the forecast events.

Some mitigation of the aforementioned concerns are possible without a complete overhaul of the modeling system. Evaluations suggest that forecast lead time, geographic location, as well as meteorological season and regime all impact the two-dimensional frequency histogram of observations relative to forecasts. This is particularly evident in the variability in the two-dimensional frequency histograms from the individual members of the CAPS SSEF, which were limited to 59 time periods and a single meteorological season. Thus, instead of creating a single analytic function to handle all scenarios, as was done in this study, one might create a continuum of analytic functions dependent on factors such as forecast lead time, meteorological regime, and season of the year. Preliminary examination suggests that all of these sensitivities are operative, but quantifying them will be challenging

and will serve as the basis for future investigations.

One additional source of error that has not been mentioned until now is the forecast bias of each numerical forecast model. As the forecast bias increases, the resulting probability forecasts change because of: 1) the number of forecast grid points increases; and 2) changes to the two-dimensional composites alter the fitting parameters. In the case of the NSSL-WRF evaluations, the forecast bias was largely negligible over the forecast dataset. This is because the forecast bias for the NSSL-WRF forecast periods was nearly one (Figures 3.1, 3.2, and 5.1). Unfortunately, the forecast bias was not always near unity for the members of the CAPS SSEF (e.g. Figures 5.2 and 5.3).

One approach to bias correcting the forecasts is to utilize quantile matching. This approach is achieved by computing the empirical cumulative distribution function from the training dataset for both the observations and the forecasts. Next the quantile of the phenomenon being forecast is determined from the observed distribution of the training dataset. Lastly, the forecast threshold associated with this quantile is computed from the forecast empirical cumulative distribution function. This computed forecast threshold is then used as the forecast threshold for the bias-corrected forecasts created from the forecast training dataset. In theory, this assures that the same fraction of forecasts occurring above the threshold and below the threshold is equal to the fraction of observations occurring above the threshold to the observations occurring below the threshold.

For example, in the case of the NSSL-WRF, if one wanted to produce a bias-corrected forecast of 12.7 mm in 6 hr using the quantile matching approach, one must find the quantile associated with that threshold from the observed empirical cumulative distribution. As illustrated in Figure 3.2, this quantile value is 0.987. The forecast threshold that corresponds to a quantile of 0.987 is 13.75 mm in 6 hr. Thus, when using quantile matching as a means of bias-correcting NSSL-WRF forecasts, a forecast accumulation of 13.75 mm or greater in 6 hr equates to observed accumulations of 12.7 mm in 6 hr or greater.

The quantile replacement approach to bias correction was used to try and improve the

forecasts from both the NSSL-WRF and the CAPS SSEF. In both cases, the bias corrected forecasts offered little, if any, improvement to the reliability — and in the case of the NSSL-WRF made the forecasts less reliable. This is due to the empirical cumulative distribution function for the training data being statistically different than for the forecast data. When the forecast and training empirical cumulative distribution functions are statistically different, the bias-corrected threshold from the training period will not be the same as the bias-corrected threshold from the forecast period. This results in "bias-corrected" probabilistic forecasts that are potentially of worse quality than non-bias-corrected forecasts.

For example, the forecast bias for the CAPS SSEF members was highly dependent on the specific time periods chosen for the training and forecast periods. This is attributed to the occurrence of a handful of time periods with significant precipitation events in which a substantial number of grid points exceeded the observed threshold. Depending on the exact number of these high amplitude time periods in each of the training or forecast dataset, the resulting empirical cumulative distribution functions tended toward one of several empirical cumulative distribution functions (Figure 5.2 and 5.3). The empirical cumulative distribution functions between the training and forecasts periods were always different making this form of bias-correction intractable.

Lastly, it is important to mention ensemble prediction systems as related to producing probabilistic forecasts. In this study, probabilistic forecasts from an ensemble prediction systems were primarily limited to probabilistic forecasts from the individual members of the ensemble. A single "ensemble probability" was produced by taking the average of all the probabilistic forecasts from the ensemble's individual members. As discussed in Section 4.2.3, the reliability of this single probabilistic forecast from the ensemble improved upon the individual members at the 25.4 mm in 6 hr threshold and was worse than the individual members at the 12.7 mm in 6 hr threshold. (See the discussion in Section 4.2.3 for more information.) Furthermore, it was shown that it is possible to produce probabilistic forecasts from each individual member that were more reliable than a

probabilistic forecast generated by post-processing the ensemble forecast (i.e., the modified Hamill and Colucci method). In light of this result, the question must be asked, "What then is the role of the ensemble?"

Traditionally, ensemble prediction systems have been used to create probabilistic forecasts. This is because well crafted ensemble prediction systems are likely to be more effective at sampling the range and character of possible solutions. As previously mentioned, KDE-based approaches rely on the underlying numerical forecast to predict the occurrence of the phenomenon being forecast. If the underlying numerical model does not forecast a phenomenon, KDE-based statistical post-processing will not produce any forecast probabilities of that phenomenon. This is where an ensemble prediction system and KDE-based statistical post-processing can compliment one another. The ensemble prediction system can provide multiple realizations from which a KDE-based statistical post-processing technique can be used. The likelihood that every member of an ensemble would miss and event is significantly less than the likelihood of any individual member missing that same event.

Taking a step back, the KDE-based post-processing technique proposed in this work has the potential to produce probabilistic forecasts from each member that are near perfect reliability (e.g., the SSEF forecasts at the 12.7 mm in 6 hr threshold used in this study). It is not readily apparent what to do with a set of fifteen probabilistic forecasts that differ between one another but are statically reliable. Each probabilistic forecast is statistically correct, but varies between one another. Should the individual member probabilistic forecasts be combined into a single probabilistic forecast? Should the probabilistic forecasts from the individual members be utilized to gain insight into the uncertainty in the spatial probabilities? In other words, how does one maximize the utility of an ensemble of probabilities? One possible solution here is to utilize both the deterministic aspect of the individual SSEF members, along with the corresponding probabilistic forecasts generated from the method proposed in this study. In this scenario, forecasters have the ability examine the evolution of the forecast in a deterministic frame work, and then utilize the corresponding probabilistic

forecasts to gain insight into the spatial uncertainty of a particular deterministic forecast.

In conclusion, the proposed method is not being offered as a replacement to ensemble prediction systems, but rather as something that can be utilized in concert with existing modeling systems. This method can be used to enhance ensemble prediction systems by providing measures of uncertainty associated with the placement of features in the individual ensemble members. At the same time, this method can produce well calibrated probabilistic forecasts from deterministic modeling systems when insufficient computing resources are available to produce a full ensemble prediction system. Ultimately, this method demonstrates that it is possible to take a model and statistically post process it in such a manner that the probabilistic forecasts are more reliable than those from an ensemble prediction system. However, this is not a guarantee that the individually post processed forecasts will beat the ensemble on any given forecast. 

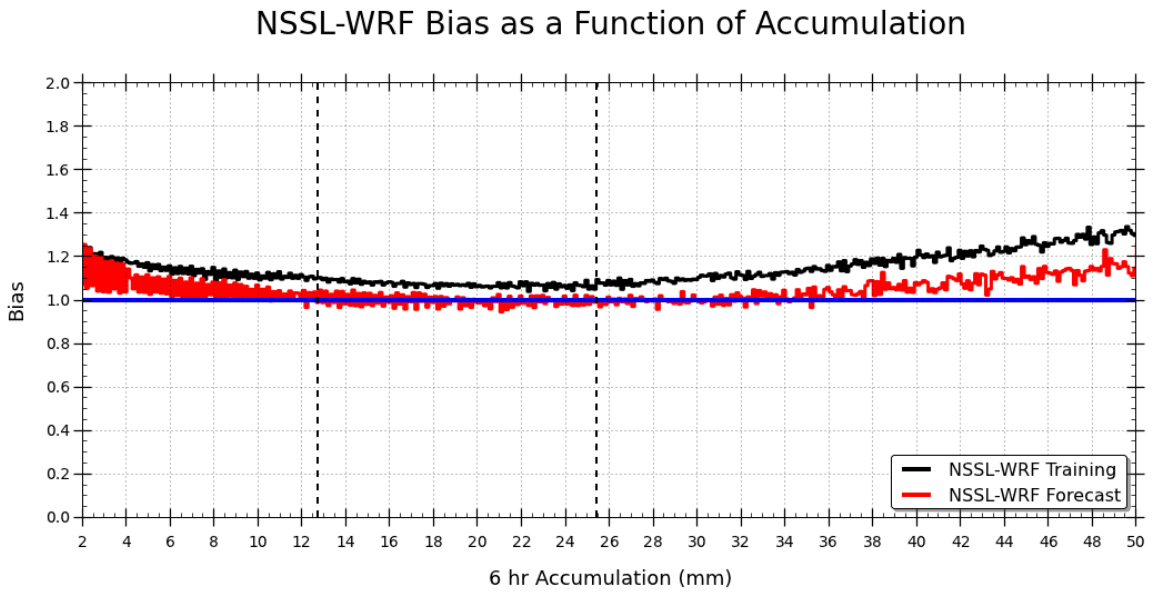

Figure 5.1: The NSSL-WRF forecast bias as a function of 6 hr precipitation threshold. The black curve is the forecast bias calculated over the 36 month training dataset. The rec curve is the forecast bias calculated over the 12 month forecast dataset. The horizontal blue line depicts the line of perfect forecast bias. The vertical black dashed lines correspond to the 12.7 mm threshold (left) and the 25.4 mm threshold (right).

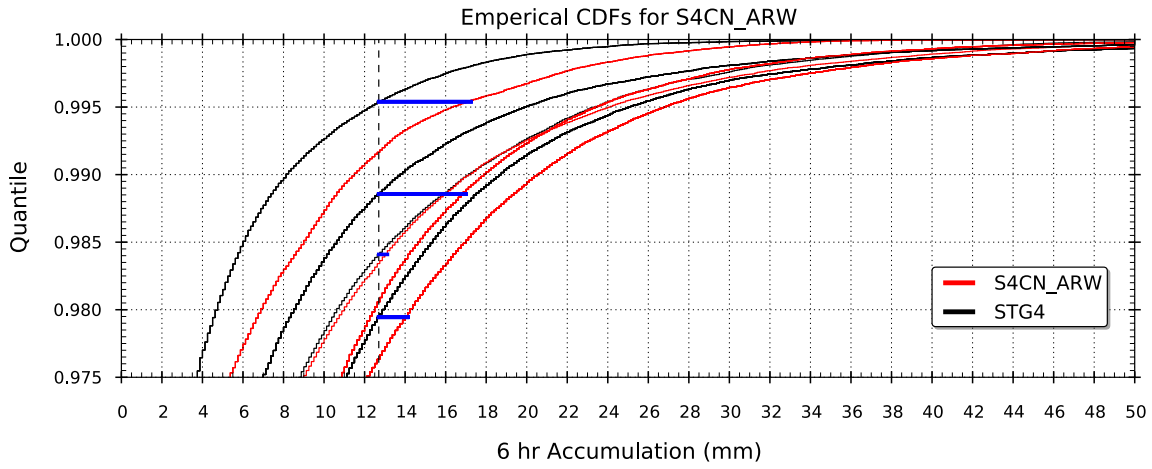

Figure 5.2: The empirical cumulative distribution functions for the twenty simulations for both the Stage IV observations (black) and the CAPS SSEF WRF-ARW control member forecasts (red). The vertical black dashed line is the 25.4 mm threshold. The horizontal blue dashed line is the connects the Stage IV quantile associated with the 12.7 mm threshold to the equivalent quantile of the WRF-ARE control member. Where the blue line intersects the WRF-ARW control member empirical cumulative distribution function is the corresponding forecast threshold at which the ratio of points above to points below is equal to the Stage IV ratio of points above to points below the 12.7 mm threshold. Note that the twenty simulations collapse into only four empirical cumulative distribution functions, depending on the dates selected.

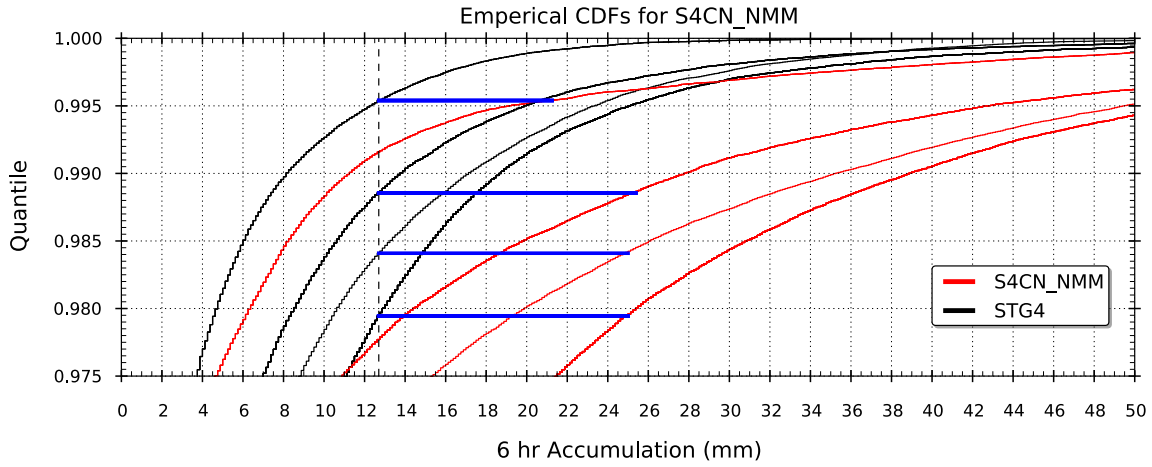

Figure 5.3: The same as in Figure 5.2, but for the CAPS SSEF WRF-NMM control member. Note the large disparity between the Stage IV empirical cumulative distribution function and the WRF-NMM control member.

References

Ångström, A., 1919: Probability and practical weather forecasting (in Swedish). *Centraltryckeriet Teknologföreningens Förlag*, 11 pp.

Ångström, A., 1922: On the effectivity of weather warnings (in Swedish). *Nord. Statistisk Tidskrift*, **1**, 394–408.

Abbe, C., 1901: The physical basis of long-range weather forecasts. *Monthly Weather Review*, **29**, 551–561.

Allen, R. L., 2001: MRF-based MOS precipitation type guidance for the United States. NWS Technical Procedures Bulletin No. 485, National Oceanic and Atmospheric Administration, U.S. Department of Commerce, 12 pp.

Allen, R. L. and M. C. Erickson, 2001: AVN-based MOS precipitation type guidance for the United States. NWS Technical Procedures Bulletin No. 476, National Oceanic and Atmospheric Administration, U.S. Department of Commerce, 9 pp.

Antolik, M. S. and C. A. Doswell III, 1989: On the contribution to model-forecast vertical motion from quasi-geostrophic processes. *Preprints, 12th Conference on Weather Analysis and Forecasting*, American Meteorological Society, Monterey, California, USA, 312–318.

Baldwin, M. E., J. S. Kain, and S. Lakshmivarahan, 2005: Development of an automated classification procedure for rainfall systems. *Monthly Weather Review*, **133**, 844–862.

Bjerknes, V., 1904: Das problem der wettervorhersage, betrachtet vom standpunkte der mechanik und der physi (the problem of weather prediction, considered from the viewpoints of mechanics and physics). *Meteorologische Zeitschrift*, **21**, 1–7, (translated and edited by Volken E. and S. Bronnimann. – Meteorologische Zeitschrift **18** (2009), 663-667).

Brooks, H. E., 1992: Operational implications of the sensitivity of modelled thunderstorms to thermal perturbations. *Preprints, 4th Workshop on Operational Meteorology*, Atmospheric Environment Service/Canadian Meteorological and Oceanographic Society, Whistler, British Columbia, Canada, 398–407.

Brooks, H. E. and C. A. Doswell, III, 1993: New technology and numerical weather prediction: A wasted opportunity? *Weather*, **48**, 173–177.

Brooks, H. E., C. A. Doswell, III, and R. A. Maddox, 1992: On the use of mesoscale and cloud-scale models in operational forecasting. *Weather and Forecasting*, **8**, 120–132.

Brooks, H. E., M. Kay, and J. A. Hart, 1998: The NCEP North American Mesoscale modeling system: Recent changes and future plans. *Preprints, 19th Conference on Severe Local Storms*, American Meteorological Society, Minneapolis, Minnesota, USA, 552–555.

Carroll, K. L., 2005: GFS-based MOS temperature and dewpoint guidance for the United States, Puerto Rico, and the U.S. Virgin Islands. MDL Technical Procedures Bulletin No. 05-05, National Oceanic and Atmospheric Administration, U.S. Department of Commerce, 8 pp.

Charney, J., 1947: On the dynamics of long waves in a baroclinic westerly current. *Journal of Meteorology*, **4**, 135–162.

Charney, J., R. Fjørtoft, and J. von Neumann, 1950: Numerical integration of the barotropic vorticity equation. *Tellus*, **2**, 237–254.

Clark, A. J., W. A. Gallus, M. Xue, and F. Kong, 2009: A comparison of precipitation forecast skill between small convection-allowing and large convection-parameterizing ensembles. *Weather and Forecasting*, **24**, 1121–1140.

Clark, A. J., W. A. Gallus, M. Xue, and F. Kong, 2010: Convection-allowing and convection-parameterizing ensemble forecasts of a mesoscale convective vortex and associated severe weather environment. *Weather and Forecasting*, **25**, 1052–1081.

Clayton, H. H., 1889: Verification of weather forecasts. *American Meteorological Journal*, **6**, 211–219.

Cooke, W. E., 1906: Forecasts and verifications in western Australia. *Monthly Weather Review*, **34**, 23–24.

Dalton, J., 1793: *Meteorological Essays*. Richarson, Phillips, and Pennington, 208 pp.

Done, J., C. A. Davis, and M. Weisman, 2004: The next generation of NWP: explicit forecasts of convection using the weather research and forecasting (WRF) model. *Atmospheric Science Letters*, **5**, 110–117.

Doswell, C. A., III and D. W. Burgess, 1988: On some issues of united states tornado climatology. *Monthly Weather Review*, **116**, 495–501.

Droegemeier, K. K., 1990: Toward a science of storm-scale prediction. *Preprints, 16th Conference on Severe Local Storms*, American Meteorological Society, Kananaskis Park, Alberta, Canada, 256–262.

Epstein, E., 1969: Stochastic dynamic prediction. *Tellus*, **21**, 739–759.

Glahn, B., M. Peroutka, J. Wiedenfeld, J. Wagner, G. Zylstra, B. Schuknecht, and B. Jackson, 2009: MOS uncertainty estimates in an ensemble framework. *Monthly Weather Review*, **137**, 246–268.

Glahn, H. R. and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *Journal of Applied Meteorology*, **11**, 1203–1211.

Gleeson, T. A., 1961: A statistical theory of meteorological measurements and predictions. *Journal of Meteorology*, **18**, 192–198.

Gleeson, T. A., 1970: Statistical-dynamic predictions. *Journal of Applied Meteorology*, **9**, 334–344.

Hallenbeck, C., 1920: Forecasting precipitation in percentages of probability. *Monthly Weather Review*, **48**, 645–647.

Hamill, T. M. and S. J. Colucci, 1998: Evaluation of EtaâĂŞRSM ensemble probabilistic precipitation forecasts. *Monthly Weather Review*, **126**, 711–724.

Hamill, T. M. and J. S. Whitaker, 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: theory and application. *Monthly Weather Review*, **134**, 3209–3229.

Hinkelmann, K., 1951: Der Mechanismus des meteorologischen Lärmes. *Tellus*, **3**, 285–296, translation: The mechanism of meteorological noise. NCAR/TN-203+STR, National Center for Atmospheric Research, Boulder, 1983.

Kain, J. S., S. R. Dembek, S. J. Weiss, J. L. Case, J. J. Levit, and R. A. Sobash, 2010: Extracting unique information from high-resolution forecast models: monitoring delected fields and phenomena every time step. *Weather and Forecasting*, **25**, 1536–1542.

Kain, J. S., et al., 2008: Some practical considerations regarding horizontal resolution in the first generation of operational convection-allowing NWP. *Weather and Forecasting*, **23**, 931–952.

Lakshmanan, V. and J. S. Kain, 2010: A gaussian mixture model approach to forecast verification. *Weather and Forecasting*, **25**, 908–920.

Leith, C., 1974: Theoretical skill of Monte Carlo forecasts. *Monthly Weather Review*, **102**, 409–418.

Lewis, J. M., 2005: Roots of ensemble forecasting. *Monthly Weather Review*, **133**, 1865–1885.

Liljas, E. and A. H. Murphy, 1994: Anders Ångström and his early papers on probability forecasting and the use/value of weather forecasts. *Bulletin of the American Meteorological Society*, **75**, 1227–1236.

Lin, Y. and K. Mitchell, 2005: The NCEP stage II/IV hourly precipitation analyses: development and applications. *Preprints, 19th Conference on Hydrology*, American Meteorological Society, San Diego, California, USA, 2–5.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *Journal of Atmospheric Sciences*, **20**, 130–148.

Lorenz, E. N., 1965: A study of the predictability of a 28-variable atmospheric model. *Tellus*, **17**, 321–333.

Lorenz, E. N., 1968: Climatic determinism. *Meteorological Monographs*, **8**, 1–3.

Lynch, P., 2008: The origins of computer weather prediction and climate modeling. *Journal of Computational Physics*, **227**, 3431–3444.

Mascart, J., 1922: The accuracy of forecasts. *Monthly Weather Review*, **50**, 592.

McPherson, R. D., 1991: 2011 – An NMC odyssey. *Preprints, 9th Conference on Numerical Weather Prediction*, American Meteorological Society, Denver, Colorado, USA, 1–4.

Murphy, A. H., 1991: Probabilities, odds, and forecasts of rare events. *Weather and Forecasting*, **6**, 302–307.

Murphy, A. H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather and Forecasting*, **8**, 281–293.

Murphy, A. H., 1996: The Finley affair: A signal event in the history of forecast verification. *Weather and Forecasting*, **11**, 3–20.

Murphy, A. H., 1998: The early history of probability forecasts: Some extensions and clarifications. *Weather and Forecasting*, **13**, 5–15.

Nichols, W. S., 1890: The mathematical elements in the estimation of the Signal Service reports. *American Meteorological Journal*, **6**, 386–392.

Olsen, D. A., N. A. Junker, and B. Korty, 1995: Ealuation of 33 years of quantitative precipitation forecasting at the NMC. *Weather and Forecasting*, **10**, 498–511.

Ortega, K. L., T. M. Smith, K. L. Manross, A. G. Kolodziej, K. A. Scharfenberg, A. Witt, and J. J. Gourley, 2009: The severe hazards analysis and verification experiment. *Bulletin of the American Meteorological Society*, **90**, 1519–1530.

Peirce, C. S., 1884: The numerical measure of the success of predictions. *Science*, **4**, 453–454.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, **133**, 1155–1174.

Roebber, P. J., 2009: Visualizing multiple measure of forecast quality. *Weather and Forecasting*, **24**, 601–608.

Rogers, E., et al., 2009: The NCEP North American Mesoscale modeling system: Recent changes and future plans. *Preprints, 23rd Conference on Weather Analysis and Forecasting/19th Conference on Numerical Weather Prediction*, American Meteorological Society, Omaha, Nebraska, USA, 363–365.

Scott, R. H., 1873/1971: On recent progress in weather knowledge. *The Royal Institution Library of Science, Earth Science*, S. K. Runcorn, Ed., Applied Science Publishers, Vol. 2, 378–387.

Sfanos, B., 2001: AVN-based MOS wind-guidance for the United States and Puerto Rico. NWS Technical Procedures Bulletin No. 475, National Oceanic and Atmospheric Administration, U.S. Department of Commerce, 6 pp.

Silverman, B. W., 1986: *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, 175 pp.

Skamarock, W. C., et al., 2008: A description of the advanced research WRF version 3. Tech. Rep. NCAR/TN–475+STR, National Center for Atmospheric Research, 113 pp., Boulder, Colorado. URL `http://www.mmm.ucar.edu/wrf/users/docs/arw_v3.pdf`.

Sobash, R. A., J. S. Kain, D. R. Bright, A. R. Dean, M. C. Coniglio, and S. J. Weiss, 2011: Probabilistic forecast guidance for severe thunderstorms based on the identification of extreme phenomena in convection-allowing model forecasts. *Weather and Forecasting*, **26**, 714–728.

Sun, W. Y. and C. Z. Chang, 1986: Diffusion model for a convective layer. Part I: Numerical simulation of convective boundary layer. *Journal of Climate and Applied Meteorology*, **25**, 1445–1453.

Theis, S. E., A. Hense, and U. Damrath, 2005: Probabilistic precipitation forecasts from a deterministic model: a pragmatic approach. *Meteorological Applications*, **12**, 257–268.

Trapp, R. J., D. M. Wheatley, N. T. Atkins, R. W. Przybylinski, and R. Wolf, 2006: Buyer beware: some words of caution on the use of severe wind reports in postevent assessment and research. *Weather and Forecasting*, **21**, 408–415.

Weisman, M. L., C. Davis, W. Wang, K. W. Manning, and J. B. Klemp, 2008: Experiences with 0–36-h explicit convective forecasts with the WRF-ARW model. *Weather and Forecasting*, **23**, 407–437.

Weiss, S. J., J. A. Hart, and P. R. Janish, 2002: An examination of severe thunderstorm wind report climatology: 1970–1999. *Preprints, 21st Conference on Severe Local Storms*, American Meteorological Society, San Antonio, Texas, USA, 446–449.

Whitnah, D. R., 1961: *A History of the United States Weather Bureau*. University of Illinois Press, 267 pp.

Wilks, D. S., 2002: Smoothing forecast ensembles with fitted probability distributions. *Quarterly journal of the royal meteorological society*, **128**, 2821–2836.

Winkler, R. L. and A. H. Murphy, 1968: "Good" probability assessors. *Journal of Applied Meteorology*, **7**, 751–758.

Word Meteorological Organization, 1992: Current trends and achievements in limited area models for numerical weather prediction research. WMO/TD No. 510, Programme on Weather Prediction Research Report Series No. 3.

Xue, M., D. Wang, J. Gao, K. Brewster, and K. K. Droegemeier, 2003: The Advanced Regional Prediction System (ARPS), storm-scale numerical weather prediction and data assimilation. *Meteorology and Atmospheric Physics*, **82**, 139–170.

Xue, M., J. Zong, and K. K. Droegemeier, 1996: Parameterization of PBL turbulence in a multi-scale non-hydrostatic model. *Preprints, 11th AMS Conference on Numerical Weather Prediction*, American Meteorological Society, Norfolk, Virginia, 363–365.